UNIVERSITÉ DE NANTES

IAE NANTES
ÉCONOMIE & MANAGEMENT

Econométrie des big data

Crude oil predictibility

Laurianne MORICEAU
Riwan PERRON

Enseignant : Mr. Darne
Année universitaire : 2020-2021

Master 2 Économétrie et Statistiques, parcours Économétrie Appliquée

Résumé

L'objectif de cette étude consiste à étudier les facteurs influençant les variations des rentabilités du pétrole. La littérature économique a montré qu'un grand nombre d'élements divers et variés permettait d'expliquer ces variations, dont certains appartenants à la sphère financière et d'autres à la sphère économique. La base de données dont nous disposions était donc composée d'un grand nombre de variables. C'est pourquoi nous traitons ici les diverses méthodes de sélection de variables ainsi que les régressions pénalisées afin de limiter les biais liés à la présence potentielle de colinéarité. La parcimonie étant de rigueur nous avons par la suite comparé les divers modèles : GETS, Ridge, Lasso, Elastic-net, Bridge, Weighted fusion, SCAD, ainsi que les méthodes adaptives. Les résultats ont montré que le modèle GETS était celui qui minimisait les critères AIC et BIC avec la sélection de 14 variables sur 33. Nous avons donc utilisé ces résultats pour modéliser un modèle par MCO qui s'est avéré robuste et dont la qualité d'ajustement est de 0.95. La modélisation a permis de conclure à l'importance de l'indice MSCI Emerging market, SP GSCI, CRB food, des rentabilités logarithmiques de l'actif Dycom Industries, du taux de change de la Nouvelle-Zélande, ainsi que de la hausse ou la baisse du prix du pétrole *ex ante* à 2 et 9 mois, dans la formation des rentabilités logarithmiques du prix du pétrole.

The main goal of this study is to analyze the factors which influence the variations of oil's rentability. Economic's literature showed that a great number of diverse and various elements could explain theses variations, and some of them belonged to the financial's sphere and the others to the economic's sphere. The data base that we had at our disposal was indeed composed of a great deal of this variables. This is why we are dealing here with the diverses selection methods of theses variables, thus the penalized regressions in order to reduce the bias linked to the potential prsence of colinearities. Parsimony being the rule, afterwards we compared the various models : GETS, Ridge, Lasso, Elastic-net, Bridge,Weighted fusion, SCAD, and also the adaptative methods. The results have shown that the GETS model was the one which minimized the AIC and BIC criterias, with the selection of 14 variables out of 33. So, we used this results to modelize a model by MCO, which proved to be sturdy and which the adjustement quality is 0,95. The modelization allows to conclude to the importance of the MSCI Emergingmarket, SP GSCI, CRB food's index, of the logarithmic rentabilities of the Dycom Industries active, of the New Zealand's exchange rates, thus of the increase ou decrease of the oil price ante 2 and 9 month, in the formation of oil price's logarithmic rentabilities.

1 Introduction

En avril 2020, la demande mondiale de pétrole chutait de 30% [1] suite à la pandémie du coronavirus, tandis que le 21 avril, on annonçait que son prix était devenu négatif. Le prix du pétrole a une influence directe sur l'économie réelle, la demande de ce produit énergétique sert au développement des secteurs industriels et des transports ; on parle alors de **marché physique**. Ainsi, lorsque l'économie est en arrêt comme lors des mesures de confinement, les besoins en pétrole sont moindres et la demande baisse entraînant une chute des prix. La production ne pouvant pas être arrêtée et les capacités de stockage étant limités, l'accumulation de barils de pétrole a entraîné des prix négatifs sur les **marchés financiers**. En outre, la fin des contrats à terme a révélé une forte présence de spéculateurs sur les marchés *futures* tandis qu'il n'y a plus assez d'acheteurs sur les marchés physiques. Les prix négatifs interviennent lorsque des primes sont appliquées sur les contrats pour lever l'obligation de prise de livraison. Cette situation exceptionnelle témoigne du lien qu'il existe entre les sphères économiques et le prix du pétrole sur les marchés financiers, le prix des contrats étant la référence pour les marchés physiques. Ainsi, l'Opep a longtemps fait la pluie et le beau temps sur les prix du pétrole et sur la conjoncture économique de certains pays, notamment pour les pays en voie de développement dont les besoins énergétiques étaient très conséquents durant les années 2000. En outre, comprendre les déterminants de l'offre et de la demande a longtemps été un enjeu pour les industriels, mais également pour les états dont les politiques économiques dépendent indirectement de ces cours. En effet, la conjoncture macroéconomique dépend du pétrole, dans la mesure où son prix détermine le pouvoir d'achat et la consommation des ménages, ainsi que le taux de marge des entreprises [2]. Pour exemple, la guerre Iran-Irak de 1980 a entraîné une baisse de 15,3 de la production mondiale de pétrole [3], et ce choc d'offre négatif a entraîné une baisse persistante de la production industrielle mondiale [4] *ie.* une hausse des prix. De plus, la crise de 2008 à l'inverse, a entraîné une chute du prix du pétrole entre juin et décembre. Les chocs structurels des prix ne seraient donc pas uniquement corrélés à l'activité économique mondiale, mais également aux cours sur les marchés financiers. On comprend alors l'importance de prévoir ces variations de prix afin de pouvoir anticiper les conséquences qu'elles pourraient engendrer.

De nombreuses études existent sur le sujet, notamment sur la volatilité et les variations des prix. A ce jour, il n'existe pas de consensus entre les économistes sur l'influence de la spéculation sur les cours, même si l'interdépendance des marchés physiques et financiers a clairement été prouvée [5], ainsi que les corrélations entre produits dérivés et actifs financiers tels que les taux de change ou les actions. Comprendre ces dynamiques est donc essentiel pour appréhender le sujet de la prédictibilité étant donné les nombreux facteurs concernés. Il existe une grande diversité de modélisation du pétrole dans la littérature académique prenant en compte ou non l'influence de variables exogènes tels que le cours des substituts au pétrole, les indicateurs macroéconomiques, la consommation etc [6]. Pour compenser les études économétriques classiques, des méthodes de machine learning comme les réseaux de neurones (NN) ou les machines à vecteurs de support (SVM) sont parfois utilisés pour intégrer la compléxitude liée à la problématique du nombre important de facteur, à la non-linéarité et la volatilité (Zhao, Jianping et Yu, 2017).

1. Wakim N., *"Le pétrole a-t-il amorcé un lent déclin ?"*, Le monde, 15 octobre 2020
2. Ibidem
3. Jess N. et al, *"Comment prévoir le prix du pétrole ?"*, Insee, Juin 2013
4. Ibidem
5. Artus P. et al, *Les effets d'un prix du pétrole élevé et volatil*, Conseil d'analyse économique (CAE), 2010
6. Yang Zhao, Jianping Li, Lean Yu, *"A deep learning ensemble approach for crude oil price forecasting"*, Energy Economics, Volume 66, 2017

L'objectif de cette étude est donc d'explorer les différentes méthodes de sélections des variables explicatives car comme évoqué précedemment il existe une multitude de facteurs expliquant le prix du pétrole. Inclure un nombre trop important de variables dans nos modèle peut entrainer du sur-apprentissage et également de l'instabilité dans l'estimation des coefficients notamment si l'étude des corrélations n'a pas été faite en amont. Dans un premier temps, nous analyserons la stationnarité des différentes variables explicatives de départ et les rendrons stationnaires si besoin afin de poursuivre notre étude. Après avoir nettoyé et décrit la base, nous utiliserons plusieurs méthodes de sélection des variables explicatives. Dans une seconde partie, nous appliquerons des modèles de régression linéaire à partir des variables sélectionnées afin de tester l'efficacité des différentes techniques de sélection sur notre base de données. La régression MCO pouvant s'écrire sous une forme matricielle :

$$Y = X\beta + \epsilon$$

où :
- Y : vecteur (89×1) de la variable expliquée (stationaire)
- X : vecteur (89×P) de variables explicatives (stationaires)
- β : vecteur (P×1) de coefficients
- ϵ : vecteur (89×1) des résidus

2 Présentation de la base

Notre étude portera sur une base de données composée de 89 observations mensuelles. Nous chercherons à expliquer la cotation du West Texas Intermédiaire (WTI), à l'aide de 34 variables explicatives. Ces variables peuvent-être réparties en 5 catégories distinctes, représentant [7] :
- Les marchés boursiers
- Les taux d'intérêt et les politiques monétaires
- L'activité économique
- Le marché des matières premières
- La confiance et l'attention des investisseurs

On cherche à expliquer les rentabilités logarithmiques des prix futures du pétrole jusqu'à décembre 2019, tel que :

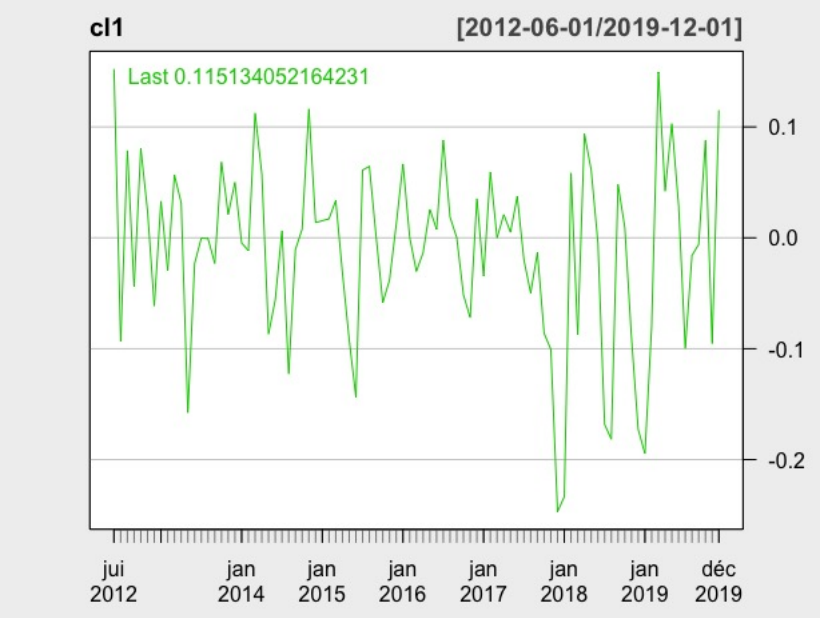
$$Rt = \frac{p_t - p_{t-1}}{p_{t-1}}$$

2.1 Etude de la stationnarité

Dans cette partie, nous allons nous intéresser au comportement de chaque variable. La figure ci-dessous montre les variations de prix du pétrole. Nous pouvons voir sur ce graphique que les rentabilités semblent stationnaires en moyenne mais pas en variance. En effet, les variations de rentabilités supposent une potentielle présence d'hétéroscédasticité.

7. Bonnier J-B., *"Explaining and forecasting cruse oil prices : An agnostic approach"* (appendix)

FIGURE 1 – Rentabilités des contrats futures



Ainsi, la série est stationnaire lorsque l'espérance mathématique et la variance sont stables dans le temps tel que : $E(y_t) = \mu \forall t \in T$, $V(y_t) = \sigma$ et $Cov(y_t, y_t^2) = \gamma_t$. Ainsi, si ces conditions sont réunies les résidus suivent un bruit blanc et la racine unitaire n'existe pas. Le test de Dickey-Fuller augmenté (ADF) permet de tester la stationnarité d'une série [8]. L'hypothèse nulle du test ADF est donc l'existence d'une racine unitaire. Pour se faire, on pose pour chaque variable explicative Xi la modélisation suivante :

$$\Delta X_{i,t} = a_i + \beta it \gamma_i X_{i,t-1} + \delta_1 \Delta X_{i,t-1} + ... + \delta_{p-1} \Delta X_{i,t-p+1} + \epsilon_t$$

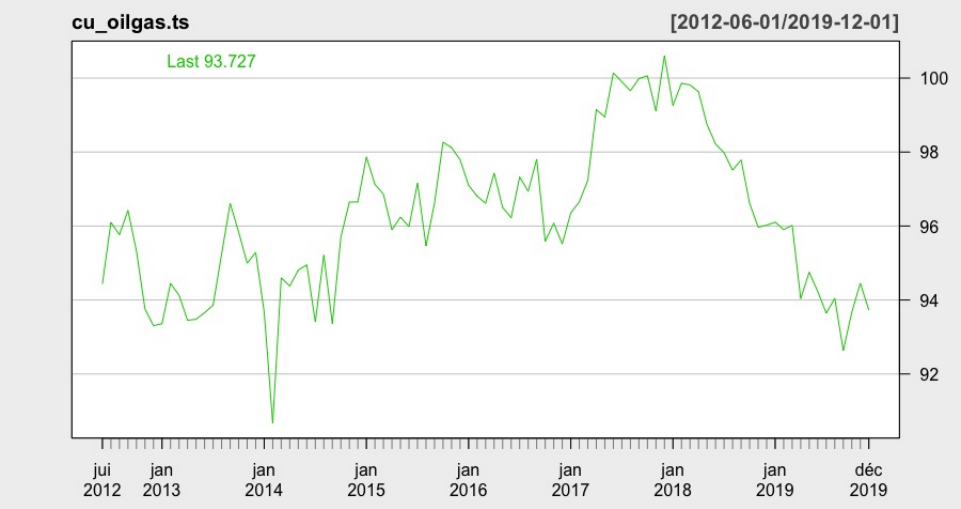
D'après Granger et Newbold (1974), la stationnarité d'une série est essentielle, lors de modèles incluant une ou plusieurs variables explicatives, il est nécessaire que ces séries soient stationnaires afin d'éviter tout risque de régression fallacieuse [9]. L'hypothèse nulle du test est que la série est non-stationnaire soit i = 0. Dans ce cas il existe une racine unitaire. L'hypothèse alternative est que i < 0, la série Xi est stationnaire, il n'existe donc pas de racine unitaire. Le nombre de retard p correspond au nombre de différenciations du modèle que nous déterminons à l'aide du critère AIC. La statistique de Dickey-Fuller nous est donnée ainsi : DF= iσ_i. Avec σ_i l'écart type du coefficient i estimé par MCO. Si la statistique est inférieure à la valeur critique au seuil de 5% donnée dans la table de Dickey-Fuller, nous rejetons l'hypothèse nulle d'existence d'une racine unitaire et acceptons l'hypothèse alternative ; la série est stationnaire. Il existe 3 types de test ADF, prenant en compte ou non l'existence d'une constante et d'une tendance temporelle. Si la tendance temporelle estimée (i) n'est pas significative, il est nécessaire de réaliser un test ne prenant pas en compte cette dernière, si la constante estimée (i) n'est pas significative dans ce second test, il est nécessaire de réaliser un test sans constante. Afin de déterminer le caractère stationnaire de la série, nous interprétons le dernier

8. DICKEY, David A. et FULLER, Wayne A. Distribution of the estimators for autoregressive time series with a unit root. Journal of the American statistical association, 1979, vol. 74, no 366a, p. 427-431.

9. GRANGER, Clive WJ, NEWBOLD, Paul, et ECONOM, J. Spurious regressions in econometrics. Baltagi, Badi H. A Companion of Theoretical Econometrics, 1974, p. 557-61.

test effectué. D'après les tests ADF, il existe une seule variable non-stationnaire : **cu_oilgas**.

FIGURE 2 – Evolution du prix cu_oilgas



Nous pouvons voir sur la figure ci-dessus que la série ne semble en effet pas être stationnaire en moyenne. Cependant, le test de Zivot et al. montre que la série n'est pas stationnaire de peu en effet la valeur du test statistique est supérieure de peu à la valeur critique au seuil de 10%. C'est pourquoi nous faisons le choix de garder la variable telle qu'elle et de la considérer comme stationnaire pour la suite des modélisations afin de ne pas perdre d'informations.

TABLE 1 – Test de Zivot et al.

Test statistique :	-4.993
Critical values :	0.01= -5.57 0.05=-5.08 0.1=-4.82

3 Statistiques descriptives

Nous pouvons retrouver ci-dessous les graphiques des séries de rentabilités et des séries en niveau.

FIGURE 3 – Graphique des rendements

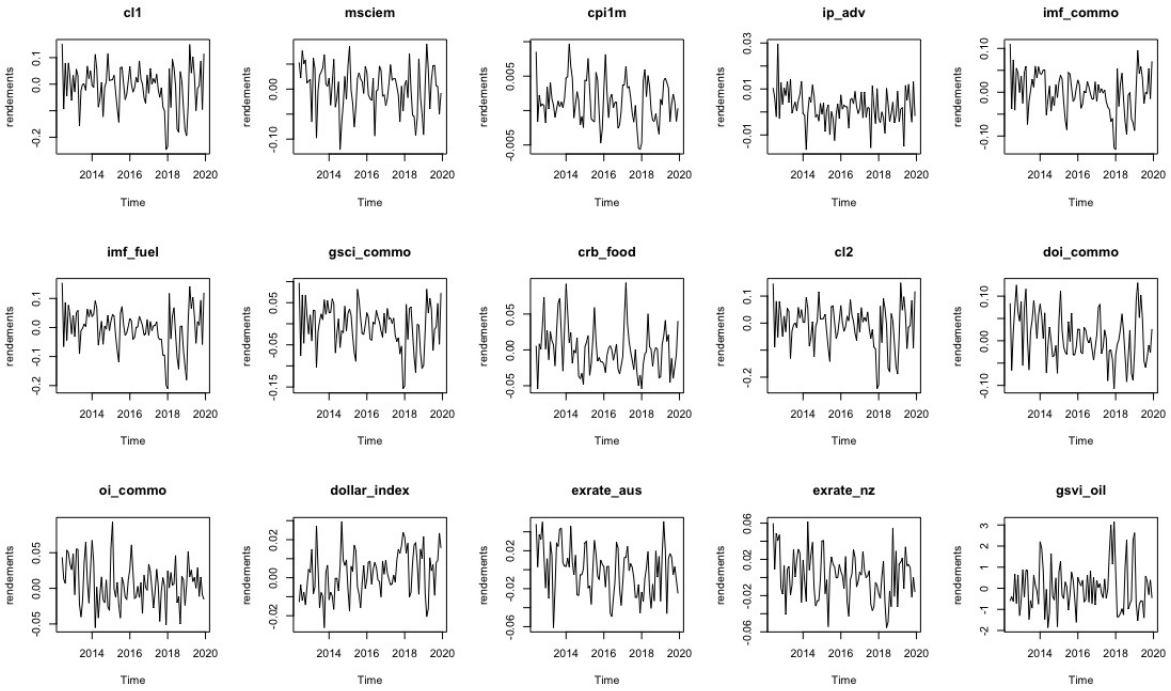
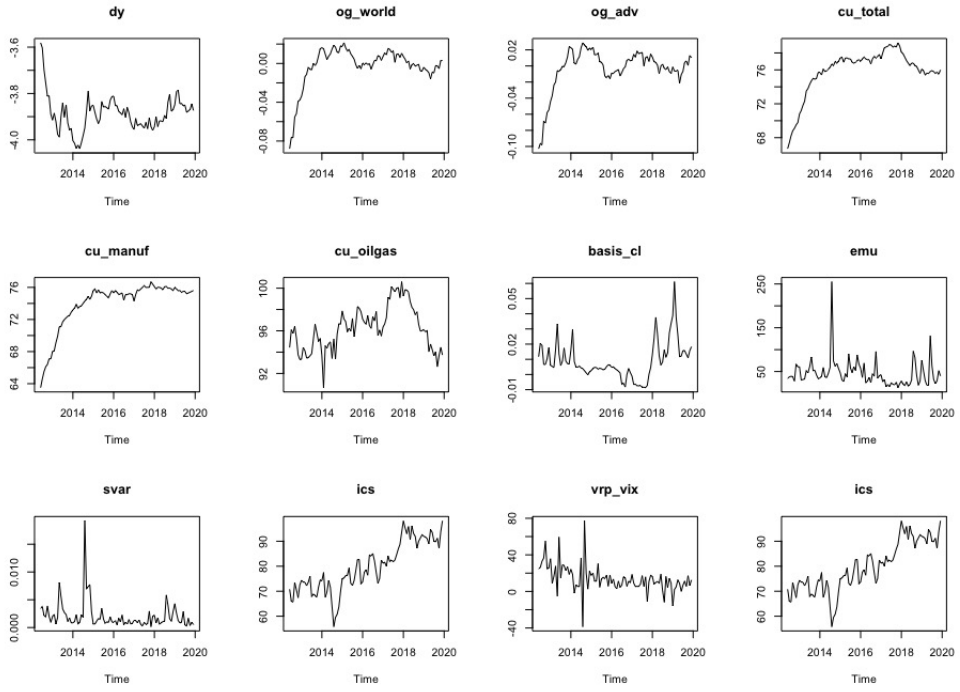


FIGURE 4 – Séries en niveau



Avant de procéder aux statistiques descriptives, il convient de nettoyer la base de données afin de savoir si des valeurs atypiques pourraient potentiellement influencer nos résultats. Nous appliquons donc la méthode de *Boudt et Al* sur les séries de rendements financiers et la méthode *Chen and Liu (1993)* sur les séries en niveau. La première méthode montre qu'il n'y a pas d'observations atypiques sur les rendements des prix du pétrole "cl1".

Nous obtenons donc les statistiques descriptives suivantes :

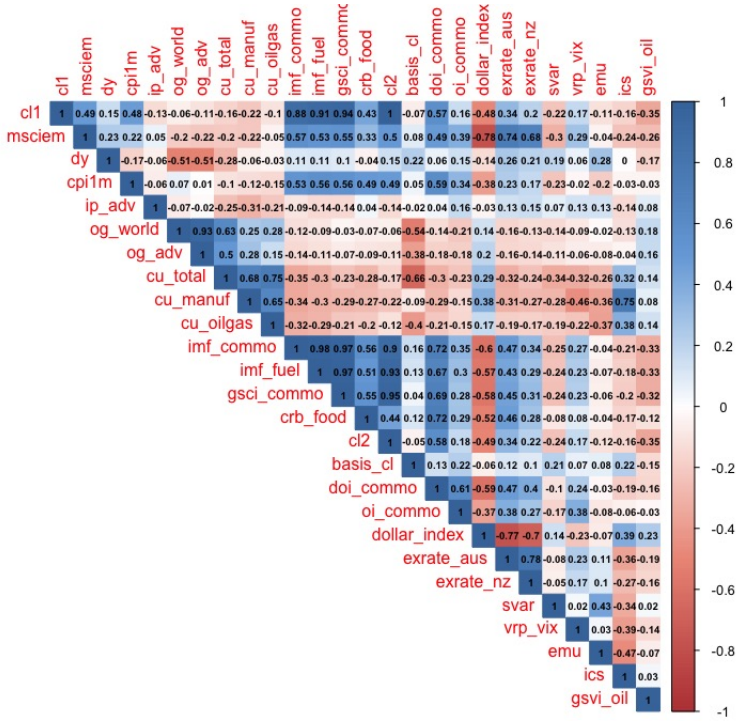
TABLE 2 – Statistiques descriptives

	nobs	Mean	Median	Variance	Stdev	Skewness	Kurtosis	AR1
cl1	91	-0.011	0.000	0.007	0.081	-0.655	0.384	0.211
msciem	91	0.002	0.008	0.002	0.044	-0.471	-0.127	0.253
dy	91	-3.884	-3.885	0.008	0.091	2.785	14.763	0.927
cpi1m	91	0.001	0.001	0.000	0.003	0.203	-0.144	0.475
ip_adv	91	0.002	0.002	0.000	0.007	-0.063	0.673	-0.092
og_world	91	-0.002	0.001	0.000	0.020	-2.224	5.345	0.992
og_adv	91	-0.004	0.001	0.001	0.026	-2.037	4.463	0.989
cu_total	91	75.837	76.700	8.470	2.910	-2.224	5.996	0.997
cu_manuf	91	73.912	75.200	10.824	3.290	-2.524	7.434	0.997
cu_oilgas	91	96.212	96.097	4.275	2.068	0.141	-0.490	0.866
imf_commo	91	0.000	-0.001	0.002	0.046	-0.404	0.138	0.299
imf_fuel	91	-0.001	0.004	0.005	0.069	-0.617	0.853	0.259
gsci_commo	91	-0.006	0.001	0.003	0.051	-0.570	0.112	0.211
crb_food	91	0.001	0.000	0.001	0.032	0.709	0.449	0.463
cl2	91	-0.009	0.004	0.006	0.078	-0.684	0.499	0.227
basis_cl	91	0.009	0.006	0.000	0.011	0.928	0.732	0.808
doi_commo	91	0.009	0.002	0.003	0.054	0.174	-0.595	0.369
oi_commo	91	0.007	0.007	0.001	0.030	0.254	-0.128	0.201
dollar_index	91	0.002	0.003	0.000	0.012	0.076	-0.536	0.401
rate_aus ex	91	0.000	0.003	0.001	0.025	-0.078	-0.602	0.349
rate_nz	91	0.002	0.002	0.001	0.026	0.056	-0.441	0.186
svar	91	0.002	0.001	0.000	0.002	1.926	3.366	0.392
vrp_vix	91	13.735	11.299	199.122	14.111	0.461	3.536	-0.056
emu	91	44.204	38.335	520.420	22.813	1.213	1.609	0.275
ics g	91	79.700	78.300	98.655	9.933	0.006	-0.865	0.935
gsvi_oil	91	0.616	-0.006	1.164	1.079	0.770	0.383	0.201

3.1 Corrélations

3.1.1 Corrélations avec la variable à expliquer

FIGURE 5 – Corrélations



Le graphique ci-dessus nous donne un apperçu des corrélations des variables explicatives avec la variable à expliquer. La multicolinéarité qui peut exister entre nos variables peut biaiser nos estimations. En effet, celle-ci peut augmenter la variance de nos coefficients, rendre certaines variables non significatives ou encore avoir des signes contradictoires avec l'effet attendu. Connaître les colinéarités permet donc de mieux anticiper des résultats instables ou invalides.

On remarque que la variable cl1 est parfaitement corrélée à la variable cl2 ce qui tombe sous le sens étant donné que la première représente le prix attendu du pétrole tandis que la deuxième représente le prix du pétrole ajusté au taux de rendement sans risque. Ainsi, nous avons décidé d'exclure la deuxième variable. On note également que la variable cl1 est très corrélée à imf_fuel et gsci_commo.

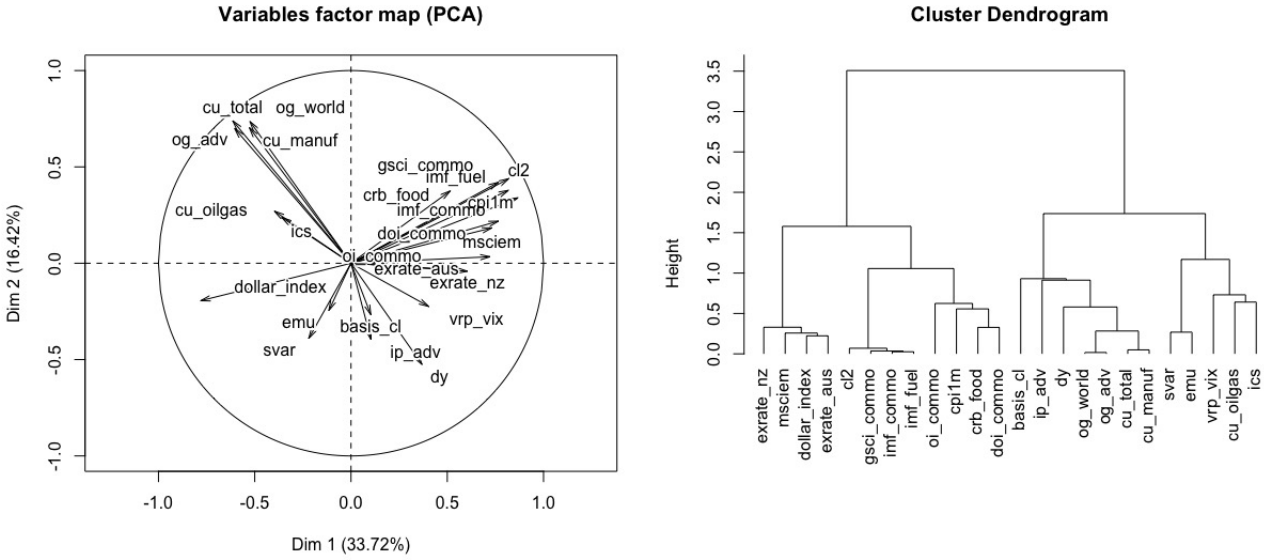
3.1.2 Corrélations entre les variables explicatives

Les variables suivantes sont également fortement corrélées entre elles :
- og_world avec og_adv
- imf_commo avec imf_fuel et gsci_commo
- imf_fuel avec gsci_commo
- gsci_commo avec cl2

3.2 Classification non supervisée

Afin de mieux comprendre comment sont liées nos variables entre elles, nous utilisons la classification par partitionnement k-means ou centres mobiles qui permet de les répartir en k classes qui permettront de minimiser de l'intertie intra-classe. L'algorithme à partir de k centres de classes choisis aléatoirement affecte à chaque individu le centre le plus proche afin d'obtenir la partition. Avec la méthode de McQueen les centres de chaque classe sont reaffectés jusqu'à ce qu'ils soient stables.

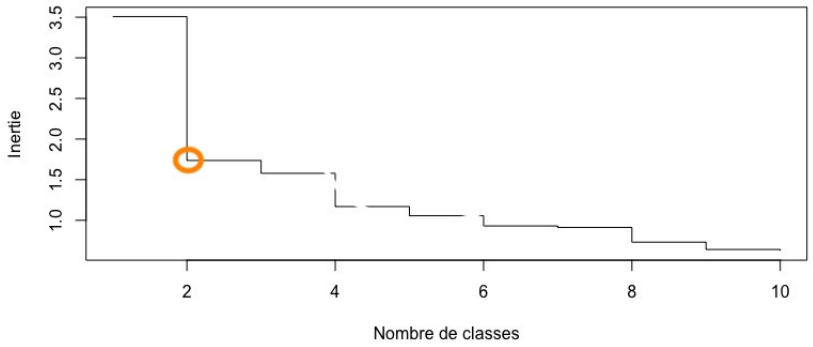
FIGURE 6 – ACP et Dendogramme des variables explicatives



Tout d'abord l'ACP avec les dimensions 1 et 2 nous permet de voir les groupes de variables quantitatives qui sont corrélées et anti-corrélées. Nous pouvons tout d'abord constater un premier groupe de variable dans le premier cadran, composé des variables **cu** qui représentent le pourcentage des ressources utilisées pour produire des biens et les variables **og** qui représentent les filtres HP utilisés dans les séries logarithmique des productions industriels pour le monde (og_world) et les pays développés (og_adv). Ce groupe pourrait donc représenter l'activité économique comme évoqué dans la présentation de la base. Il s'oppose par l'axe 1 aux variables **basis_cl**, **ip_adv** et **dy** qui représentent la sphère financière. Dans le cadran en haut à droite nous retrouvons les variables représentant le marché des matières pemières. Ces variables s'opposent aux variables **dollar_index**, **emu** et **svar** et donc aux variables monétaires.

Cette partition ne nous donne pas de classification stricte, c'est pourquoi nous avons également fait une classification hierarchique sur nos variables, représenté à droite. Le graphique ci-dessous représente l'inertie de notre densogramme. Il montre qu'il y a un 'gap' au niveau de la classe 2 entouré ici par un cercle orange. Cette classification en deux classes montre une opposition entre les variables financières et les variables économiques.

FIGURE 7 – Inertie du dendogramme



4 Modèles de selection

Il existe plusieurs modèles de sélection de variables : les méthodes économétrique et les procédés statistiques de Machine Learning. L'objectif de cette partie est d'obtenir une sélection de variables explicatives à partir de chaque méthode que nous comparerons. Certaines méthodes peuvent avoir tendance à choisir de manière trop restrictive le nombre de variable alors que d'autres au contraire montrent des difficultés à choisir et ne diminuent pas assez la dimension des variables explicatives.

4.1 Approche éconmétrique (GETS)

La sélection de variables est un problème que se posent les statisticiens depuis de nombreuses années. En effet, nous pourrions tester toutes les combinaisons de variables possibles mais cette méthode serait trop gourmande en puissance algorithmique, car le nombre de régression linéaires à effectuer sera alors de 2^P si l'on souhaite explorer toutes les combinaisons possibles. Les méthodes Forward, Backward et Stepwise qui reposent sur l'ajout et/ou le retrait successif de variables dans une régression de type MCO, visent à maximiser le Log-Vraissemblance et à minimser les critère de Aikaike et Bayséens afin d'optimiser le nombre de « chemins » à tester pour trouver un nombre de combinaisons de variables satisfaisantes. Cependant, le choix d'une variable selectionnée à un certain point dépend des choix précédents effectués et sont donc dépendant du chemin emprunté pour arriver à la solution, on parle alors de *path dependency*.

De multiples méthodes ont été développées pour automatiser cette sélection. Nous allons dans un premier temps nous pencher sur l'approche **GETS (General to Specific)**, basée sur la méthode Backward. L'idée ici est créer des chemins multiples (Multi-Path) à l'aide retraits successifs de variables sous certaines conditions en partant du modèle général. Notre série de départ étant hautement stationnaire, nous décidons de réaliser un premier modèle non restreint général (General Unrestricted Model) sans prendre en compte de partie autorégressive endogène retardée en plus des variables explicatives. Nous vérifions alors l'autocorrélation et l'hétéroscédasticité des résidus avec les tests de Ljung-Box et Box-Pierce, ainsi que sur la normalité avec le test de Jarque-Bera.

Les résultats des tests Ljung-Box nous suggèrent qu'il n'y a pas d'autocorrélation ni d'hétéroscédasticité dans la distribution des résidus. Le test de Jarque-Bera montre également que les résidus semblent distribués normalement. Nous décidons de garder cette modélisation pour l'approche GETS. Ce modèle a donc été intégré dans la modélisation du GETS sans prise en compte de l'hétéroscédasticité. L'algorithme va retirer successivement des variables une par une à partir d'un seuil de significativité $\alpha = 5\%$ en vérifiant une série de test diagnostique sur les résidus. L'ordre par lequel sont retirées les variables est appelé *'chemin'* que l'algorithme emprunte, ici 2. Le modèle optimal est choisi parmi les chemins empruntés par l'algorithme selon le critère de Schwarz (SC). Une seule des variables **exrate_nz** n'est pas significative au seuil de 5%, ce qui signifie qu'elle a été réintroduite uniquement pour respecter les tests de diagnostic.

TABLE 3 – Résultats du modèle GETS

	coef	std.error	t-stat	p-value	
msciem	0.1746627	0.0587823	2.9713	0.0039541	**
dy	0.0043985	0.0012125	3.6277	0.0005120	***
ip_adv	0.6392190	0.2077172	3.0774	0.0028922	**
imf_commo	-0.9001756	0.1537906	-5.8533	1.115e-07	***
imf_fuel	0.6649373	0.1264720	5.2576	1.269e-06	***
gsci_commo	1.4167260	0.1637914	8.6496	5.685e-13	***
crb_food	-0.2108547	0.0525696	-4.0110	0.0001389	***
oi_commo	-0.1703218	0.0721597	-2.3603	0.0207902	*
exrate_nz	-0.2588322	0.0935498	-2.7668	0.0070854	**
svar	2.3597449	0.8731454	2.7026	0.0084612	**
ma19_cl	-0.0055509	0.0034119	-1.6269	0.1078381	
mom2_cl	0.0220095	0.0040698	5.4080	6.929e-07	***
mom9_cl	0.0147036	0.0040045	3.6718	0.0004423	***
mom12_cl	-0.0096325	0.0036159	-2.6639	0.0094037	**

Diagnostiques

	Chi-sq	df	p-value
Ljung-Box AR(1)	2.39427	1	0.1218
Ljung-Box ARCH(1)	0.62017	1	0.4310

SE of regression	0.01729
R-squared	0.96115
Log-lik (n=91)	247.14286

Note : 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

4.2 Sélection de variables à partir de régressions pénalisées

Lorsque l'on traite de grande bases de données avec un nombre de variables explicatives important, il peut arriver d'obtenir une variance élevée des estimateurs MCO entraînant des coefficients instables et donc très dépendants de l'échantillon d'apprentissage. En effet, la variance dépend de la matrice des coefficients des régresseurs X'X qui doit être inversible. La variance estimée de l'erreur étant égale à $\sigma_\epsilon^2 = \frac{\sum_{i=1}^{n} \epsilon_i^2}{n-p}$, lorsque le nombre de variable p est bien supérieur au nombre d'observations, alors la matrice XX' n'est pas inversible.

De plus, on peut voir apparaître des problèmes de colinéarité liée à la dépendance des variables explicatives entre elles, entraînant également une instabilité dans l'estimation des coefficients.

Ainsi, les régressions penalisées viennent pallier à ces différents problème en appliquant **le principe de régulation** qui consiste à accepter une augmentation du biais afin d'obtenir une réduction de la variance en imposant des contraintes sur les coefficients afin d'encadrer l'amplitude de leurs valeurs. On parle alors de *shrinkage*, tel que :

$$\min_{\beta_1,\dots,\beta_p} \sum_{i=1}^{n} (y_i - \sum_{j=1}^{p} \beta_j z_{ij})^2$$

sous contrainte de pénalité $\sum_{j=1}^{p} \beta_j^q \leq \tau$

Où les variables x_i sont centrées réduites et y_i est au minimum centrée pour évacuer la constante.

De manière équivalente en intégrant la fonction de pénalité nous obtenons :

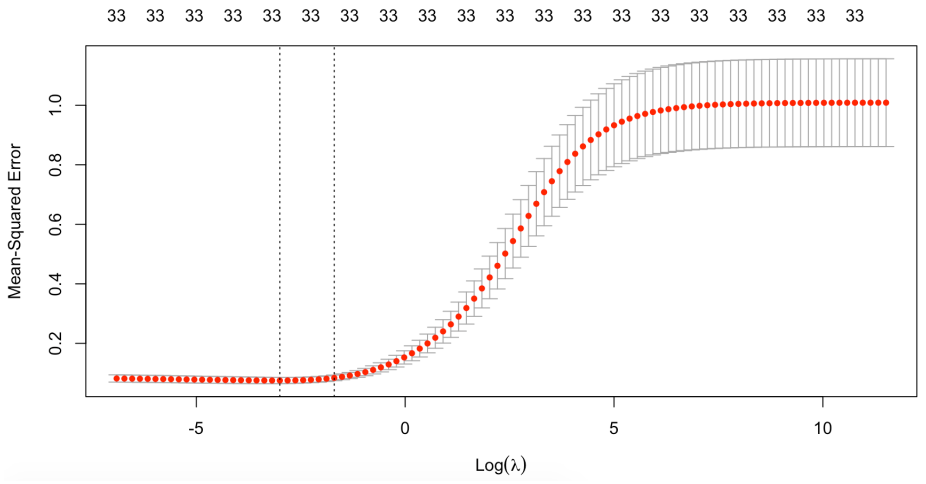
$$\min_{\beta_1,\dots,\beta_p} \sum_{i=1}^{n} (y_i - \sum_{j=1}^{p} \beta_j z_{ij})^2 + \lambda \sum_{j=1}^{p} \beta_j^q$$

Où λ est le coefficient de pénalité. Lorsque $q = 1 \Rightarrow \beta_j$, nous utiliserons la régression **LASSO**. Lorsque $q = 2 \Rightarrow \beta_j^2$, nous utiliserons la régression **Ridge**

4.2.1 La régression Ridge

La régression Ridge, proposée par Hoerl et Kennard (1970), consiste au rétrécissement de l'amplitude des estimateurs vers zéro, et donc de leur influence sur la variable dépendante à l'aide d'une régression pénalisée. Par définition, les estimateurs de Ridge sont définis même si la matrice des estimateurs n'est pas de rang plein et donc non-réversible. Les estimateurs sont également de variance plus faible que ceux d'une régression MCO, cependant ils comportent un biais dans leur estimation du fait de la pénalisation appliquée. Le paramètre λ contrôle la 'puissance' de rétrécissement des estimateurs vers 0. Il augmentera le poids de la pénalité affecté à chaque coefficient. La régression Ridge est donc très utile pour un nombre important de variable avec une grande probabilité de colinéarité. Nous évaluons donc la méthode en prenant une plage de cent valeurs de lambda allant de 10^{-3} à 10^5. L'erreur moyenne la plus faible peut être observée dans le graphique ci-dessous :

FIGURE 8 – Erreur moyenne au carré par lambda - Ridge



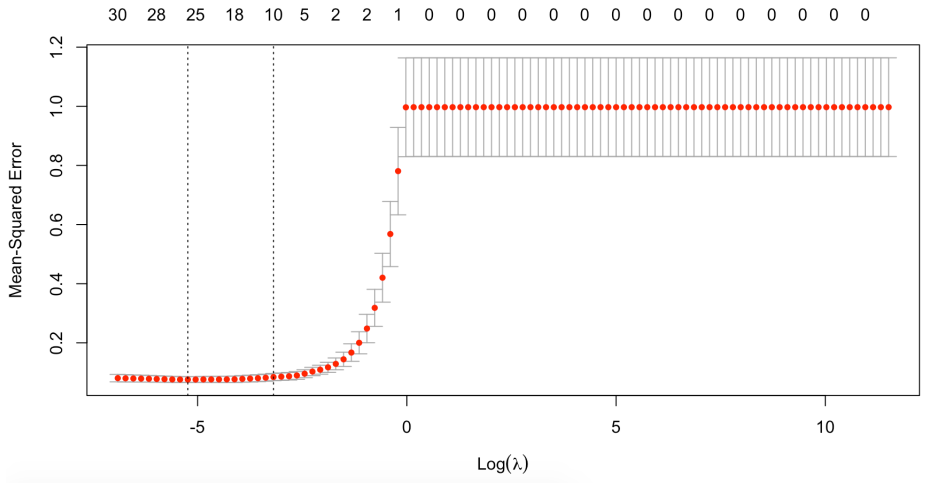
Nous pouvons voir que toutes les variables ont été conservées, en effet, la méthode Ridge n'est pas capable de sélectionner les variables, mais seulement de diminuer de manière très forte les coefficients associés aux variables ayant le moins d'impact sur la variable à expliquer. La fonction *cv.glmnet* permet de sélectionner le pramètre λ optimal. Ici, le lambda nous donnant l'erreur moyenne d'estimation la plus faible en validation croisée est : $\lambda = 0{,}0498$. On obtient alors un vecteur de coefficients β_i pour chaque λ_i. En outre, les valeurs absolues des coefficients permettent de hierarchiser les variables les plus importantes de la régression. Nous obtenons également un taux d'erreur minimal en validation croisée de 0.004.

En conclusion, l'approche Ridge permet de résoudre les problèmes de multi-colinéarité tout en conservant l'intégralité des variable explicatives d'origine. Nous allons voir maintenant comment la méthode LASSO permet de réduire la dimension de nos variables explicatives.

4.2.2 Régression LASSO

La méthode de *Least Absolute Shrinkage and Selection Operator* ou LASSO est une méthode de sélection de variable par régression pénalisée proposée par Tibshirani (1996). L'objectif cette fois-ci consiste à éliminer les variables inutiles. Son fonctionnement est proche de celui de Ridge, la particularité réside en l'algorithme d'optimisation convexe qui peut provoquer une pénalité sur l'estimation des coefficients conduisant à la mise à 0 d'un certain nombre de ces derniers. L'algorithme LARS peut être utilisé dans l'estimation des coefficients et la mise à 0 de certain. Cette solution présente l'avantage d'être parcimonieuse par rapport à la méthode Ridge. En revanche, lorsque qu'il existe une colinéarité élevée entre certaines variables, l'algorithme risque d'éliminer arbitrairement une variable importante dans l'explication de Y (Zou et Hastie, 2005). De plus, si (P > N), l'algorithme LASSO ne sélectionnera mécaniquement que N variables prédictives au maximum. Nous évaluons donc la méthode en prenant une plage de cent valeurs de lambda allant de 10^{-3} à 10^{5}. L'erreur moyenne la plus faible peut être observée dans le graphique ci-dessous :

FIGURE 9 – Erreur moyenne au carré par lambda - LASSO

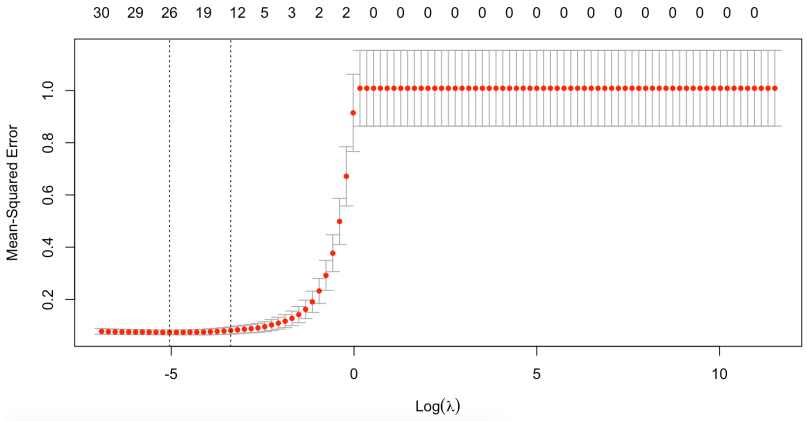


Pour choisir le lamba, on prend le lambda pour lequel l'erreur moyenne est minimisée et augmenté de sont écart-type. Les résultats montrent qu'à partir de $\lambda = 0{,}0413$ le taux d'erreur n'est plus réduit de manière significative. La méthode sélectionne donc 7 variables pour ce lambda avec la méthode LASSO.

4.2.3 Elastic-Net

La régression Elastic-Net vient en complément de la régression LASSO afin de corriger les possibles erreurs liées à la colinéarité qui est traitée dans la régression Ridge. Ainsi la régression Elastic-Net ajoute une pénalisation Ridge à celle de LASSO. Proposée par Zhou et Hastie (2005), elle permet comme dans la régression Ridge de moins discriminer les variables trop corrélées qui étaient néanmoins importante dans la conception du modèle final en partageant les poids des groupes de variables prédictives corrélées. Tout comme la régression LASSO, les variables non pertinentes ont un coefficient nul et sont écartées pour la construction des modèles (sélection de variables). Il a un effet de regroupement *'grouping effect'* s'il existe un ensemble de variables parmi lesquelles les corrélations par paire sont élevées, alors Elastic-Net regroupe les variables corrélées. L'idée est ici de faire varier le paramètre α entre 0 et 1 afin de donner plus de poids à l'une ou l'autre des pénalisations de Ridge ou de LASSO. Ceci, afin de calculer ce paramètre α optimum. Nous définierons une plage de alpha à explorer, pour lesquelles nous effectuerons une validation croisée sur la plage des cent valeurs de lambda allant de 10^{-3} à 10^{5}. L'objectif ici est de déterminer les α et λ optimaux conjointement. Nous avons donc inclu α dans une boucle afin de comparer les taux d'erreur.

FIGURE 10 – Erreur moyenne au carré par lambda - Elastic Net

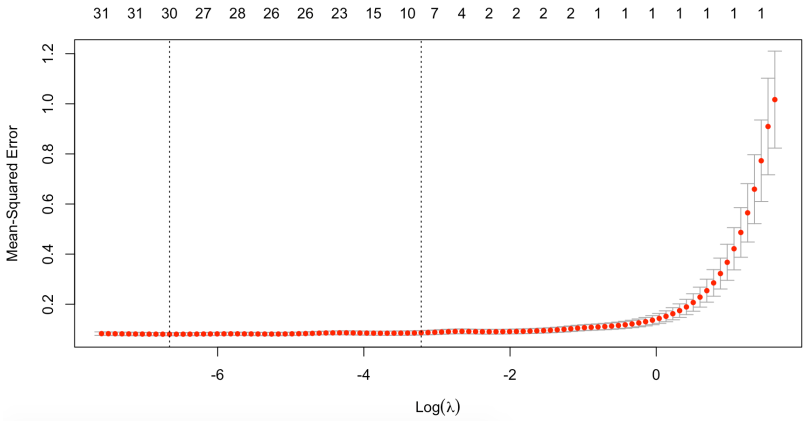


Pour choisir le lamba, on prend le lambda pour lequel l'erreur moyenne est minimisée et augmenté de sont écart-type. Les résultats montrent qu'à partir de $\lambda = 0{,}0343$ le taux d'erreur n'est plus réduit de manière significative. La méthode sélectionne donc 10 variables pour ce lambda pour un $\alpha = 0.15$ avec la méthode Elastic-Net. Nous pouvons voir que l'algorithme a bien selectionné 12 des variables que nous retrouverons dans le tableau récapitulatif.

4.2.4 Adaptive LASSO

Afin de résoudre le problème de sélection des variables du LASSO, Zhou (2006) a proposé l'Adaptive LASSO (aLASSO). La méthode consiste en l'inclusion d'une première pondération des variables dans un modèle LASSE afin d'aider l'algorithme en lui donnant tout d'abord une information sur le poids des différentes variables. aLASSO utilise donc le même système de pénalité que le LASSO avec l'inclusion d'un paramètre de pondération qui provient d'un modèle réalisé en amont qui peut être LASSO, Ridge mais également un modèle OLS. Dans notre exemple, nous utiliserons un premier modèle LASSO afin de pondérer les poids des variables puis nous intégrerons ces poids dans un nouveau modèle avec une pénalité factorielle par pondération de la racine carré des poids. Le graphique de l'erreur moyenne en fonction de la force de la pénalité peut être observée dans le graphique ci-dessous :

FIGURE 11 – Erreur moyenne au carré par lambda - adaptive LASSO



Pour choisir le lamba, on prend le lambda pour lequel l'erreur moyenne est minimisée et augmenté de sont écart-type. Les résultats montrent qu'à partir de $\lambda = 0{,}040$ le taux d'erreur n'est plus réduit de manière significative. La méthode sélectionne donc 10 variables pour ce lambda avec la méthode adaptive-LASSO.

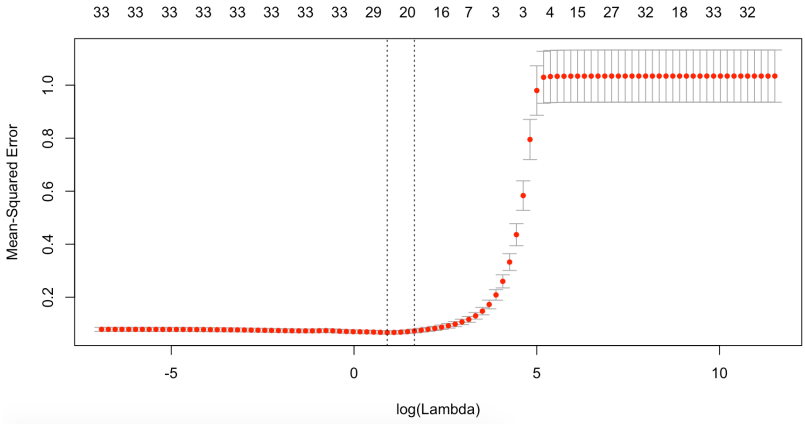
4.2.5 Bridge

La régression Bridge proposée par Frank et Firedman (1993), généralise les régressions Ridge et LASSO en utilisant la norme Lq, c'est-à-dire le type de pénalisation effectuée :

$$\min_{\beta_1,\ldots,\beta_p} \sum_{i=1}^{n}(y_i - \sum_{j=1}^{p}\beta_j z_{ij})^2 + \lambda \sum_{j=1}^{p}\beta_j^q$$

De la même manière que pour le Lasso ou le Bridge, on cherche à minimiser la somme des carrés des résidus pénalisés pour chaque $\lambda \geq 0$. Le paramètre q permet d'encadrer la plage de contrainte. Le graphique de l'erreur moyenne en fonction de la force de la pénalité (log(lambda)) peut être observée dans le graphique ci-dessous :

FIGURE 12 – Erreur moyenne au carré par lambda - Bridge



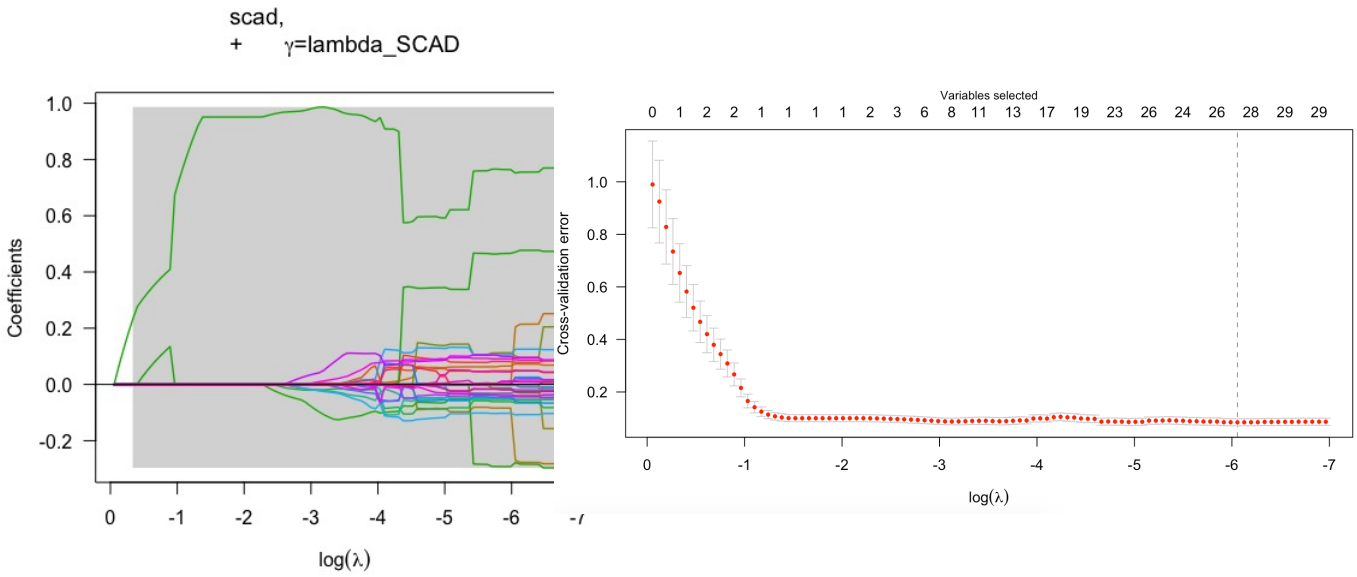
Pour choisir le lamba, on prend le lambda pour lequel l'erreur moyenne est minimisée et augmenté de sont écart-type. Les résultats montrent qu'à partir de $\lambda = 5{,}214$ le taux d'erreur n'est plus réduit de manière significative. La méthode sélectionne donc 20 variables pour ce lambda avec la méthode Bridge.

4.2.6 SCAD

La méthode SCAD (Smooth Cut-Off Absolute Deviation) est une alternative à la méthode LASSO proposée par Fan et Li (2001) et Zhang (2010). Le problème avec la méthode LASSO réside dans le fait que lorsque le nombre d'observations est important et le nombre de variables explicatives est faible, la puissance de la régression (λ) se doit d'être très forte si l'on souhaite annihiler l'effet des variables parasites. Cela implique un biais très important dans l'estimation des coefficients de la régression pénalisée. Le principe de la méthode SCAD pour résoudre ce problème est d'appliquer des puissances de pénalités différentes en fonction du coefficient estimé. Si le coefficient estimé par LASSO est élevé, la pénalisation sera limitée pour

minimiser le biais dans l'estimation. En revanche, pour les valeurs de beta faibles la pénalité sera au contraire plus importante. Le graphique de l'erreur moyenne en fonction de la force de la pénalité peut être observée dans le graphique ci-dessous :

FIGURE 13 – Chemins de régulations et erreur moyenne

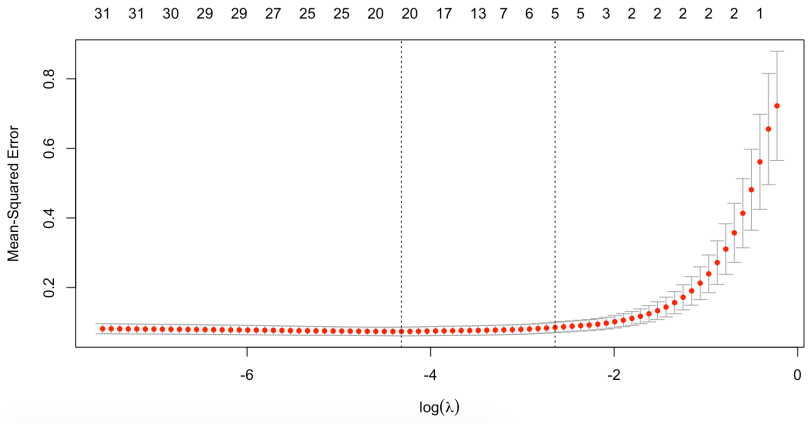


Pour choisir le lamba, on prend le lambda pour lequel l'erreur moyenne est minimisée. Les résultats montrent qu'à partir de $\lambda = 0{,}014$ le taux d'erreurest minimisé. La méthode sélectionne donc 27 variables pour ce lambda avec la méthode SCAD.

4.2.7 Weighted fusion

Daye et Jeng (2009) proposent la méthode Weighted fusion pour les données corrélées et utiles lorsque p > N. Le graphique de l'erreur moyenne en fonction de la force de la pénalité peut être observée dans le graphique ci-dessous :

FIGURE 14 – Erreur moyenne au carré par lambda - Weighted Fusion



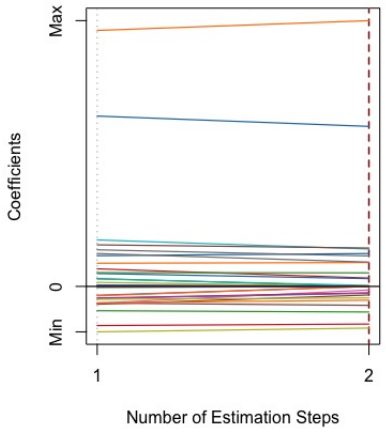
Pour choisir le lamba, on prend le lambda pour lequel l'erreur moyenne est minimisée . Les résultats montrent qu'à partir de $\lambda = 0{,}013$ le taux d'erreur est minimisé. La méthode sélectionne donc 20 variables pour ce lambda avec la méthode Weighted Fusion.

4.2.8 Méthodes adaptives (Xiao & Xu)

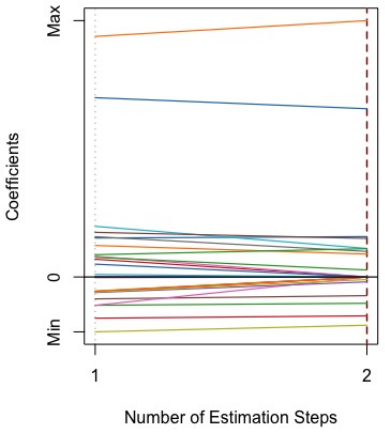
Proposé par Xiao et Xu (2015), le multi-step adaptive EN est un modèle itératif dans lequel on utilise des paramètres de régulations utilisés séparement à chaque étape afin de réduire le nombre de faux positifs (variables) afin de maintenir une estimation précise. L'algorithme par 'multi-step' permet de savoir quelles sont les variables éliminées à chaque étape et ainsi avoir une plus grande connaissance des corrélations entre les groupes de variables [10]. Les figures ci-dessous représentent l'effet relatif de chaque variable sur les estimations. On peut observer l'évolution du critère d'information afin de selectionner l'étape optimale. Le trait orange représente la variable 12 : gsci_commo et le trait bleu la variable 11 : imf_fuel.

10. Xiao N., Xu Q-S., *"Multi-step adaptive elastic-net : reducing false positives in high-dimensional variable selection*, Journal of Statistical Computation and Simulation, Volume 85, 2015

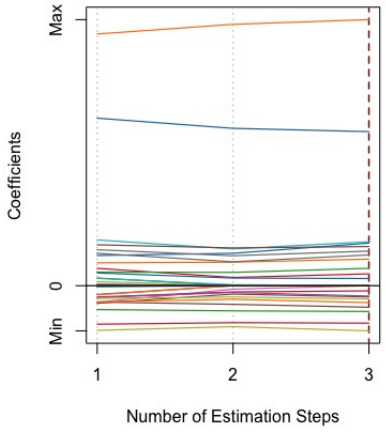
FIGURE 15 – Estimations par méthodes adaptives



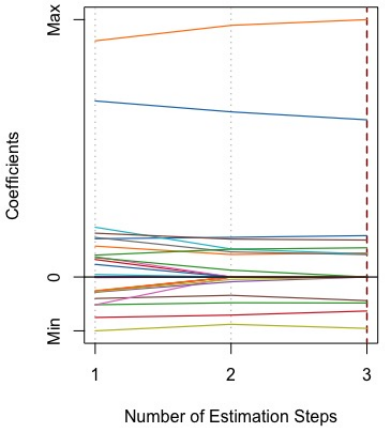
(a) Adaptive EN



(b) Adaptive SCAD



(c) MS-aEN



(d) MS-aSCAD

4.2.9 Approche complémentaire : Smooth Lasso

Proposé par Hebiri et Van de Geer (2010), la régression smooth lasso propose une pénalisation adaptée lorsque les coefficients de régressions successifs varient lentement ou pour les variables ordonnées [11], l'objectif consiste toujours à minimiser l'erreur moyenne tel que :

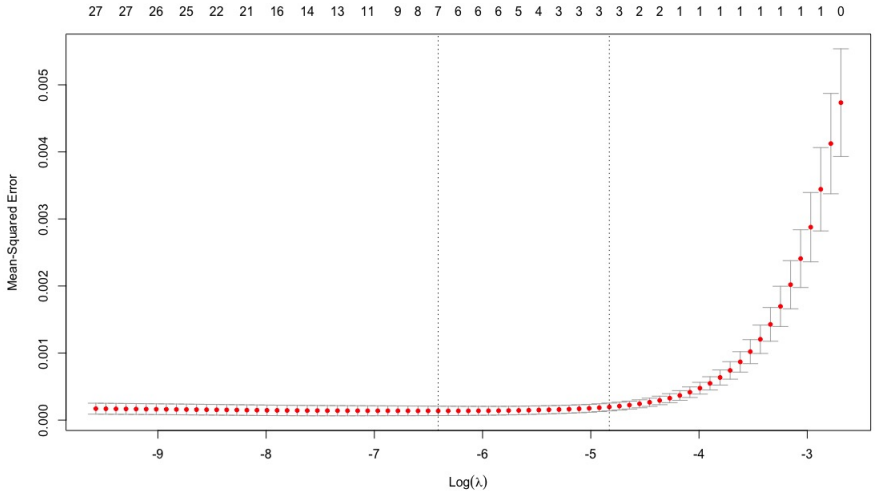
$$\beta_{SLasso} = \min_{\beta_1,\ldots,\beta_p} \frac{1}{2n} \sum_{i=1}^{n} (y_i - \sum_{j=1}^{p} \beta_j z_{ij})^2 + \lambda \sum_{j=1}^{p} \beta_j^q + \mu \sum_{j=2}^{p} (\beta_j - \beta_{j-1})^2$$

11. El Anbari Mohammed, Régularisation et sélection de variables par le biais de vraisemblance pénalisée, HAL

Où λ et μ sont les paramètres qui contrôle le lissage des estimateurs.

En outre, cet estimateur permet de mieux interpréter les modèles car il localise les endroits où les coefficients sont égaux à zéro tandis que les autres coefficients sont estimés avec une régulation permise grâce à deux fonctions de penalités.

FIGURE 16 – Erreur moyenne au carré par lambda - Smooth Lasso



5 Récapitulatif et comparaison des résultats

Le tableau ci-dessous résume les choix de variables pour chaque modèle. Nous observons dans ce tableau que la méthode adaptive LASSO est la plus parcimonieuse avec la prise en compte de seulement 6 variables explicatives. On peut également remarquer que certaines variables font partie de tous les modèles : imf_fuel et $gsci\_commo$. Nous pouvons donc suggérer que ces variables donnent une information importante sur les variations de notre variable dépendante y. On rappelle que les méthodes Ridge et Bridge ne sélectionnent pas de variable mais affectent des poids différents à chaque variable. La méthode économétrique 'General To Specific' propose un entre deux avec la sélection de la moitié des variables. Notons également que le modèle aLASSO est très restrictif avec seulement 6 variables selectionnées, au contraire de SCAD ou aEN qui en gardent plus d'une vingtaine.

FIGURE 17 – Liste des varibles prises en compte dans chaque modèle

Variables / modèles	Gets	Ridge	LASSO	EN	aLASSO	Bridge	WF	SCAD	aEN	aSCAD	MS-aEN	MS-aSCAD	S-LASSO	Total
msciem	1	1	0	1	0	1	1	1	1	1	1	1	1	11
dy	1	1	1	1	0	1	1	1	1	1	1	1	1	12
cpiOUIm	0	1	0	0	0	1	0	0	0	0	0	0	0	2
ip_adv	1	1	0	0	0	1	0	1	1	0	1	0	0	6
og_world	0	1	0	0	0	1	0	1	0	0	0	0	0	3
og_adv	0	1	0	0	0	1	0	0	1	0	1	0	0	4
cu_total	0	1	0	0	0	1	0	0	0	1	0	0	0	3
cu_manuf	0	1	0	0	0	1	0	1	1	0	1	0	0	5
cu_oilgas	0	1	1	1	0	1	1	0	0	0	0	0	1	6
imf_commo	1	1	0	0	0	1	1	1	0	0	0	0	0	5
imf_fuel	1	1	1	1	1	1	1	1	1	1	1	1	1	13
gsci_commo	1	1	1	1	1	1	1	1	1	1	1	1	1	13
crb_food	1	1	1	1	0	1	1	1	1	1	1	1	1	12
basis_cl	0	1	1	1	1	1	1	1	1	1	1	1	1	12
doi_commo	0	1	0	0	0	1	0	1	0	0	0	0	0	3
oi_commo	1	1	1	1	0	1	1	1	1	1	1	1	1	12
dollar_index	0	1	0	0	0	1	0	1	1	0	0	0	0	4
exrate_aus	0	1	0	0	0	1	0	0	0	0	0	0	0	2
exrate_nz	1	1	1	1	0	1	1	1	1	1	1	1	1	12
svar	1	1	0	1	0	1	1	1	1	1	1	1	1	11
vrp_vix	0	1	0	0	0	1	0	0	0	0	0	0	0	2
emu	0	1	0	0	0	1	1	1	1	0	1	0	1	7
ics	0	1	0	0	0	1	0	0	0	1	0	0	0	3
gsvi_oil	0	1	1	1	1	1	1	1	1	0	1	0	1	10
maOUI9_cl	1	1	0	0	0	1	0	1	1	1	1	0	0	7
mom2_cl	1	1	1	1	1	1	1	1	1	1	1	1	1	13
mom3_cl	0	1	0	0	0	1	0	0	0	0	0	0	0	2
mom9_cl	1	1	1	1	0	1	1	1	1	1	1	1	1	12
momOUI2_cl	1	1	0	1	0	1	1	1	1	1	1	0	1	10
volOUI9_cl	0	1	1	1	0	1	1	0	1	0	0	0	1	7
vol29_cl	0	1	0	1	0	1	1	1	1	0	1	0	1	8
vol39_cl	0	1	0	0	0	1	1	1	0	0	0	0	0	4
vol2OUI2_cl	0	1	1	1	1	1	1	0	1	1	1	1	1	11
Total	14	33	13	17	6	33	20	23	22	16	20	12	18	

Nous allons maintenant établir des modèles de régression linéaire par la méthode des moindres carrés ordinaires. L'ajustement sera réalisé sur notre échantillon de données. Il ne sera pas pris en compte de partie autoregressive AR(1) car comme montré à l'aide de la méthode GETS et évoqué dans la littérature existante, les prix retardés d'une période du pétrole ne donnent pas d'information sur les prix du pétrole présent.

6 Régression par MCO

Nous réalisons donc 13 modèles différents en prenant en compte pour chaque méthode de séléction les variables séléctionnées dans la partie précédente. La figure 18 représente une liste de tests et d'indicateurs sur l'ajustement de nos modèles et les résidus de chaque modèle. Nous nous interesserons plus particulièrement à la qualité d'ajustement à l'aide du coefficient de détermination qui représente la variance des prix du pétrole expliquée par chaque modèle par rapport à la variance totale des prix du pétrole sur notre échantillon. L'erreur moyenne nous donnera également une information sur la qualité d'ajustement du modèle. Nous chercherons également à savoir quel modèle maximise la fonction de vraissemblance (Log-Likehood) qui traduit la capacité du modèle à retranscrite la fonction de distribution de notre variable à expliquer. Plus le Log-Likehood est élevé, mieux le modèle explique la distribution.

Afin de prendre en compte la parcimonie des différents modèles, nous nous interesserons aux modèles qui minimisent les critères AIC et BIC qui prennent en compte la qualité d'ajustement tout en considérant le nombre de variables explicatives du modèle. Ceci dans le but d'avoir un modèle qui ne serait pas biaisé par un sur-ajustement à l'échantillon utilisé. En effet, l'utilisation d'un des modèles dans le but de prévoir ou d'expliquer les variation des prix mensuels du pétrole ne doit pas être spécifique à notre échantillon d'entrainement. En outre, le sur-apprentissage risque d'augmenter légèrement l'erreur moyenne d'ajustement sur notre échantillon (biais) mais de diminuer la variance des estimateurs si l'on effectue l'ajustement à l'aide d'autres échantillons. Dans le cas de la prévision, l'objectif est de minimiser la variance des prédictions sur de nouveaux échantillons.

Dans le but de valider l'inférence statistique, nous effectuerons un certain nombre de tests sur les résidus, notamment la normalité avec le test de Jarque-Bera où lorsque la p-value est supérieure au seuil de 5%, l'hypothèse H0 de distribution normale est acceptée et les résidus suivent une loi normale. Nous pourrons comparer ce test avec les valeurs Skewness et Kurtosis. Nous effectuerons également les tests d'autocorrélation des erreurs LB-AR(1) et d'autocorrélation de la variance des erreurs LB-ARCH(1). L'hypothèse H0 de ces tests est qu'il n'existe pas d'autocorrélation (p-val >0,05).

FIGURE 18 – Tests statistiques sur les régressions MCO

Tests /modèles	gets	Ridge	LASSO	EN	aLASSO	Bridge	WF	SCAD	aEN	aSCAD	MSaEN	MSaSCAD	S-LASSO
Nombre Variables	14	33	13	17	6	33	20	23	22	16	20	12	18
R2	0.9616	0.9683	0.9563	0.9589	0.9383	0.9683	0.9635	0.9663	0.9645	0.9589	0.9634	0.956	0.9607
Erreur standard	0.0158	0.0144	0.0169	0.0163	0.02	0.0144	0.0154	0.0148	0.0152	0.0163	0.0154	0.0169	0.016
LogLikehood	248.2698	257.0296	242.4334	245.2181	226.6614	257.0296	250.6244	254.2195	251.8116	245.2274	250.4614	242.0903	247.2642
AIC	-464.5395	-444.0592	-454.8668	-452.4362	-437.3227	-444.0592	-457.2488	-458.439	-455.6232	-454.4548	-456.9228	-456.1806	-454.5285
BIC	-424.3658	-356.1792	-417.2039	-404.7299	-417.2359	-356.1792	-402.0099	-395.6675	-395.3625	-409.2593	-401.6839	-421.0286	-404.3113
P.val LB-AR(1)	0.1317	0.3521	0.1678	0.1628	0.1055	0.3521	0.306	0.2676	0.3561	0.2218	0.3363	0.1349	0.3248
P.val LB-ARCH(1)	0.516	0.5046	0.5432	0.6543	0.869	0.5046	0.9069	0.7301	0.3179	0.2611	0.566	0.6946	0.8579
Jarque Bera	0.3835	0.1192	0.6805	0.8416	0.0097	0.1192	0.6156	0.4334	0.4726	0.9	0.6497	0.36	0.5758
skewness	-0.1267	-0.312	0.1907	0.13	0.7223	-0.312	0.0732	-0.1764	0.0072	0.0888	-0.0127	0.2577	0.1609
Kurtosis	0.5826	0.764	0.159	0.0764	0.4619	0.764	0.4073	0.4811	0.5493	0.0827	0.4003	0.4373	0.3541

Nous observons dans ce tableau que les modèles ayant la qualité d'ajustement la plus élevée sont les modèles issus des méthodes Ridge et Bridge qui prennent en compte l'intégralité des variables explicatives (l'utilité de ces méthodes est donc limité dans notre analyse de séléction de variables). Ils possèdent le R2 et le Log-Likehood les plus élevés (respectivement 0,9683 et 257,03) ainsi que l'erreur d'ajustement la plus faible (0,014). En effet, la regression par MCO a essayé de capter l'intégralité de l'information fournie par nos variables explicative à notre variable à expliquer, bien que toutes les variables aient étés tirées d'une

analyse économique préalable, il se peut qu'une partie de l'information fournie par certaines variables ne soit pas de la "vraie" information mais bien des variations résiduelles concordantes entre ces variables et notre variable dépendante. Ces variations sont alors propres à l'échantillon d'ajustement choisi et ne se vérifierons malheureusement pas sur un échantillon d'ajustement différent.

Afin de limiter ce risque, nous devons nous efforcer d'utiliser un modèle parcimonieux qui n'utilise que les variables ayant un apport important dans l'explication de notre variable à expliquer. Nous remarquons que le modèle qui minimise de manière nette les critères AIC et BIC (respectivement -464,54 et -424,37) est le modèle utilisant les variables selectionnées à partir de la méthode économétrique GETS. Nous pouvons remarquer que la méthode GETS sélectionne 14 variables sur 33 avec une qualité d'ajustement très proche des modèles prenant en compte l'intégralité des variables avec un R2 = 0.962, un Log-Likehood = 248,27 et une erreur moyenne d'ajustement = 0,016).

Nous nous interessons donc aux résidus de ce modèle, nous observons une P-value du test de Jarque-Bera de 0,3835, ce qui signifie que l'hypothèse d'une distribution normale est respectée. La valeur de l'excès de skewness est proche de 0 ce qui signifie que les résidus sont ditribués de manière symétrique, en revanche il semblerait qu'il y ait un léger excès de kutosis, soit une distribution légèrement leptokurtique. De plus, il n'existe pas d'autocorrélation dans les erreurs, les p-value des tests LB-AR(1) et LB-ARCH(1) étant supérieurs à 0,05 (respectivement 0,13 et 0,52). Nous pouvons maintenant observer la significativité des différents coefficients estimés. Le test de significativité de Student permet de mettre en rapport la variance d'estimation et la valeur estimé de chaque coefficient afin de savoir s'il parait pertinent de l'utiliser. En d'autres termes, que la variable associée à ce coefficient a un impact significatif sur notre variable à expliquer. Les tests de Student pour chaque modèle sont récapitulés dans la figure 19.

FIGURE 19 – P-value des tests de Student sur les coefficients pour chaque modèle

Variables / Modèles	gets	Ridge	LASSO	EN	aLASSO	Bridge	WF	SCAD	aEN	aSCAD	MS-aEN	MS-aSCAD	S-LASSO
msciem	0,06241331	0,24721962		0,17315171		0,24721962	0,10927182	0,23710701	0,19990507	0,03628699	0,03917503	0,05074459	0,18340699
dy	0,04546422	0,3723094	0,0077804	0,04840784		0,3723094	0,02724713	0,1577675	0,15970919	0,07492984	0,14685432	0,00948508	0,04144638
cpi1m		0,7932997				0,7932997							
ip_adv	0,0430942	0,05819152				0,05819152		0,06073821	0,05685966		0,1001538		
og_world		0,27454522				0,27454522		0,65538996					
og_adv		0,18804814				0,18804814			0,30614371		0,3668086		
cu_total		0,41823836				0,41823836				0,98394319			
cu_manuf		0,27427272				0,27427272		0,4662779	0,09083386		0,13354804		
cu_oilgas		0,77167386	0,13082735	0,32721936		0,77167386	0,85625283						0,53661067
imf_commo	0,00010898	0,11517262				0,11517262	0,03111338	0,05180938					
imf_fuel	6,54E-06	0,00155178	0,00066841	0,00093084	0,00010425	0,00155178	0,0002141	0,0004739	0,00293064	0,00056746	0,00203991	0,00110301	0,00298278
gsci_commo	1,69E-12	1,95E-05	2,10E-06	3,40E-06	0,00035981	1,95E-05	6,81E-07	4,89E-07	4,10E-07	2,70E-06	3,93E-07	1,43E-06	9,40E-07
crb_food	0,02167682	0,14432535	0,04991125	0,0410107		0,14432535	0,08155467	0,08862849	0,04363604	0,02949593	0,06638792	0,05037318	0,05058022
basis_cl		0,0582385	0,020308	0,01480302	9,99E-05	0,0582385	0,04874703	0,07988158	0,00305205	0,00840504	0,0029018	0,00867195	0,0072405
doi_commo		0,56450764				0,56450764		0,64821832					
oi_commo	0,0112411	0,48274247	0,08981048	0,07379031		0,48274247	0,11082803	0,22329023	0,06786401	0,06256904	0,06520303	0,0302432	0,07678775
dollar_index		0,55725072				0,55725072		0,22019126	0,27647039				
exrate_aus		0,65596442				0,65596442							
exrate_nz	0,01064422	0,04263553	0,01274151	0,01063746		0,04263553	0,02405123	0,00810161	0,00352384	0,00351341	0,00530715	0,0011272	0,01621099
svar	0,64033973	0,05029834		0,21686122		0,05029834	0,12112549	0,08578108	0,00610909	0,18696659	0,007555	0,40529979	0,03424799
vrp_vix		0,78289432				0,78289432							
emu		0,03210257				0,03210257	0,17043878	0,09963624	0,0104003		0,0174159		0,07294427
ics		0,95461617				0,95461617				0,55465628			
gsvi_oil		0,50691773	0,31532829	0,40261286	0,10211487	0,50691773	0,49046553	0,48750512	0,34011014		0,42669997		0,29192233
ma19_cl	0,2846952	0,24850493				0,24850493		0,3044875	0,41105701	0,13690682	0,2490701		
mom2_cl	5,76E-05	0,09268383	0,01455154	0,01331753	0,02408112	0,09268383	0,00909064	0,00744782	0,03597891	0,00465573	0,01798955	0,01193764	0,02567674
mom3_cl		0,67980455				0,67980455							
mom9_cl	0,00114286	0,01813556	0,03359417	0,01244545		0,01813556	0,01059319	0,00392434	0,00382487	0,01415031	0,00503486	0,05513157	0,01284458
mom12_cl	0,19028701	0,29257748		0,19251075		0,29257748	0,16227786	0,23350147	0,32960302	0,37182548	0,37363013		0,19153526
vol19_cl		0,79796913	0,23870974	0,32115863		0,79796913	0,69895331		0,4367797				0,38887902
vol29_cl		0,36498892		0,37431796		0,36498892	0,30821635	0,18411809	0,55655379		0,52099433		0,43270945
vol39_cl		0,41413638				0,41413638	0,77721396	0,50166335					
vol212_cl		0,91198998	0,28364698	0,81952516	0,19795652	0,91198998	0,96534663		0,69313942	0,07492723	0,53527622	0,05206084	0,76864633

Nous pouvons dans un premier temps observer que les 2 variables sélectionnées par toutes les méthodes possèdent égalementdes coefficients significatifs dans l'ensemble des modèles réalisés : imf_fuel et $gsci\_commo$. Cela nous confirme la pertinence de ces variables pour notre modèle final. La méthode Gets sélectionne donc 14 variables explicatives, trois de ces variables ne sont pas significatives au seuil de 10% ($mom12\_cl$,$ma19\_cl$ et $svar$). Une n'est pas significative au seuil de 5% (msciem). Nous essaierons par la suite d'utiliser un modèle qui ne prend pas en compte ces variables.

Nous allons maintenant vérifier la présence de multicolinéarité entre nos variables explicatives. En effet, la multicolinéarité pourrait provoquer une instabilité des coefficients, la variance des coefficients serait alors très forte et leurs valeurs varieraient fortement en fonction de l'échantillon d'ajustement choisi, ce qui limite grandement l'interprétation de notre modèle. Nous calculons donc le Facteur d'Inflation de Variance (VIF) pour l'ensemble des variables de nos modèles (Figure 20). Nous fixons un seuil à 10, au dessus duquel il est considéré qu'il existe une multicolinéarité trop importante entre les variables.

FIGURE 20 – Variance Inflation Factor

Variables / Modèles	gets	Ridge	LASSO	EN	aLASSO	Bridge	WF	SCAD	aEN	aSCAD	MS-aEN	MS-aSCAD	S-LASSO
msciem	3,3479	6,2709		3,8224		6,2709	3,8998	4,7185	4,7777	3,5861	3,8295	3,3673	3,8257
dy	1,6232	5,0064	1,3589	1,6611		5,0064	1,7153	3,5746	3,2502	2,2115	3,1055	1,3761	1,6618
cpi1m		2,7506				2,7506							
ip_adv	1,2856	2,1979				2,1979		1,5425	1,667		1,5819		
og_world		104,511				104,511		8,9795					
og_adv		79,6261				79,6261			8,2043		7,6465		
cu_total		82,5663				82,5663				2,6655			
cu_manuf		61,2959				61,2959		7,7875	7,5255		7,1234		
cu_oilgas		5,5832	1,432	1,7048		5,5832	2,0404						1,7746
imf_commo	34,9202	70,4214				70,4214	51,4351	53,3979					
imf_fuel	27,2877	38,3553	19,9893	20,1375	17,0953	38,3553	31,7681	33,8389	20,4364	18,902	19,9352	18,2615	20,9657
gsci_commo	23,0276	49,4369	20,9472	21,3902	16,7294	49,4369	36,5593	38,1497	22,7613	21,7593	22,5857	20,7795	22,2276
crb_food	1,7712	2,8812	1,6548	1,7022		2,8812	1,7317	2,455	2,0196	1,8298	1,9219	1,6649	1,7096
basis_cl		4,1193	1,6123	1,7091	1,1929	4,1193	1,9252	2,1169	1,6355	2,2068	1,6238	1,5068	1,7429
doi_commo		7,4989				7,4989		5,939					
oi_commo	1,4462	3,0358	1,3968	1,507		3,0358	1,5443	2,6821	1,6399	1,4787	1,6043	1,4244	1,508
dollar_index		5,235				5,235		3,8672	3,9308				
exrate_aus		6,7035				6,7035							
exrate_nz	2,0957	4,2981	1,3849	2,4513		4,2981	2,4985	2,6261	2,6496	2,302	2,3848	2,2061	2,4775
svar	1,5695	8,1422		1,613		8,1422	3,1312	4,4153	4,1214	2,3152	4,1066	1,5146	2,9231
vrp_vix		2,7698				2,7698							
emu		4,9459				4,9459	2,7899	3,2885	2,9875		2,8616		2,6071
ics		9,5206				9,5206				2,6194			
gsvi_oil		1,9438	1,4493	1,5273	1,3839	1,9438	1,651	1,7368	1,6455		1,6232		1,5484
ma19_cl	1,8186	2,8832				2,8832		2,141	2,1195	1,8803	2,0017		
mom2_cl	1,7676	3,7926	1,7442	1,8534	1,6175	3,7926	2,0202	2,6204	2,2448	2,0877	2,1745	1,6719	1,8951
mom3_cl		3,2449				3,2449							
mom9_cl	1,4099	1,8998	1,4297	1,5689		1,8998	1,5829	1,6648	1,6524	1,6286	1,6434	1,3711	1,5701
mom12_cl	1,6087	2,1289		1,6939		2,1289	1,7015	1,7887	1,7958	1,7773	1,7863		1,694
vol19_cl		1,7388	1,2759	1,3082		1,7388	1,3769		1,4829				1,3164
vol29_cl		4,1271		3,3135		4,1271	3,3767	1,9971	3,4108		3,4039		3,3273
vol39_cl		2,382				2,382	1,6829	1,7319					
vol212_cl		6,0079	1,8109	3,6792	1,5039	6,0079	3,9274		3,7318	1,7928	3,5496	1,7113	3,6836

Les variables imf_commo, imf_fuel, et $gsci\_commo$ semblent présenter de la multicolinéarité. Ce qui est logique puisque ces variables étaient très corrélées dans le tableau des corrélations présenté en partie précédente. Nous partons du modèle avec les 14 variables sélectionnées par la méthode GETS afin de tester un nouveau modèle en retirant deux des trois variables possédant un VIF très élevé de manière arbitraire. Ceci, afin de limiter le risque de multicolinéarité. Nous retirons également les variables qui n'étaient pas significatives au seuil de 10%.

Les résultats de ce modèle sont les suivants :

Call : $lm(cl1 \sim msciem + dy + ip\_adv + gsci\_commo + crb\_food + oi\_commo + exrate\_nz + mom2\_cl + mom9\_cl, data = training\_dlbase)$

Table 4 – Modèle à 9 variables

Coefficients :

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	0.300231	0.123258	2.436	0.017054	*
msciem	0.119201	0.075983	1.569	0.088063	.
dy	0.080909	0.031833	2.542	0.012940	*
ip_adv	0.341251	0.316466	1.078	0.284092	
gsci_commo	1.499409	0.064713	23.170	< 2e-16	***
crb_food	-0.242083	0.082785	-2.924	0.004476	**
oi_commo	-0.252603	0.081752	-3.090	0.002744	**
exrate_nz	-0.350961	0.112267	-3.126	0.002460	**
mom2_cl	0.017472	0.005022	3.479	0.000812	***
mom9_cl	0.012839	0.004720	2.720	0.007977	**

Diagniostics :

Residual standard error :	0.01993
Multiple R-squared :	0.9457
Adjusted R-squared :	0.9396

Note : 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Table 5 – VIF - Modèle à 9 variables

msciem	dy	ip_adv	gsci_commo	crb_food	oi_commo	exrate_nz	mom2_cl	mom9_cl
2.594061	1.341495	1.193758	2.525235	1.563800	1.352739	1.930972	1.455497	1.203739

On remarque que la multicolinéarité a disparu, ce qui montre que les coefficients de ce modèle sont plus stables. Cependant, une variable auparavant significative au seuil de 5% n'est maintenant plus significative (ip_adv). Nous décidons donc de la retirer afin d'essayer d'obtenir un modèle final dont tous les coefficients sont significatifs afin de ne pas biaiser les autres coefficients. Les résultats du modèle final sont disponibles dans la figure ci-dessous :

Call : $lm(cl1 \sim msciem + dy + gsci\_commo + crb\_food + oi\_commo + exrate\_nz + mom2\_cl + mom9\_cl, data = training\_dlbase)$

Table 6 – Modèle à 8 variables

Coefficients :

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	0.313303	0.122782	2.552	0.012577	*
msciem	0.116356	0.076013	1.531	0.093159	.
dy	0.084261	0.031713	2.657	0.009473	**
gsci_commo	1.482136	0.062761	23.615	<2,00E-16	***
crb_food	-0.235771	0.082659	-2.852	0.005491	**
oi_commo	-0.233008	0.079786	-2.920	0.004513	**
exrate_nz	-0.330402	0.110746	-2.983	0.003754	**
mom2_cl	0.017834	0.005015	3.556	0.000629	***
mom9_cl	0.013465	0.004689	2.872	0.005193	**

Diagniostics :

Nombre Variables :	8
Residual standard error :	0.01993
Multiple R-squared :	0.9457
Adjusted R-squared :	0.9396
LogLikehood :	232.0322
AIC :	-444.0645
BIC :	-418.9559
P.val LB-AR(1) :	0.0609
P.val LB-ARCH(1) :	0.9025
Jarque Bera :	0.8013
Skewness :	0.1154
Kurtosis :	0.1774

Note : 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Table 7 – VIF - Modèle final à 8 variables

msciem	dy	gsci_commo	crb_food	oi_commo	exrate_nz	mom2_cl	mom9_cl
2.590934	1.328705	2.370506	1.555983	1.285902	1.875282	1.448989	1.185569

Dans ce modèle final, il n’existe aucun risque de multicolinéarité, 7 coefficients sont significatifs au seuil de 5% et 1 au seuil de 10%. Ce modèle parait très robuste et parcimonieux. La qualité d’ajustement sur l’échantillon est plus faible que les modèles GETS et Ridge avec un R2 de 0,945, un Log-Likehood de 232 et une erreur de prévision de 0,02. Le critère AIC est dans la moyenne des autres modèles (-444,06) et le critère BIC est bon et proche de celui du modèle établie à partir de la méthode GETS (-418,96). Les test sur les résidus nous montrent que l’hypothèse de normalité des résidus est largement confirmée avec une p-value Jarque-Bera = 0,80, skewness = 0,12, Kurtosis = 0,18), l’hypothèse des MCO est vérifiée. Il n’existe pas d’autocorrélation des résidus (p-value test LB-AR(1) = 0,06 et LB-ARCH(1) = 0,90).

Nous pensons que ce modèle est celui qui possède le meilleur rapport robustesse/précision par rapport

aux autres modèles. Il respecte de manière forte les hypothèses de régression formulées par les MCO, il est très parcimonieux, ne souffre pas de multicolinéarité et tous ses coefficients sont significatifs au seuil de 5% ou 10%. Nous pensons qu'il est préférable de l'utiliser lors de l'interprétation des coefficients car il offre des coefficients illustrant de manière correcte les relations entre nos variables explicatives et notre variable dépendante. La modélisation a permis de conclure à l'impact positif de l'indice MSCI EM, SP GSCI, de l'actif Dycom industries, ainsi que le prix du pétrole à 2 et 9 mois. En revanche, l'indice CRB Food, le crude oil ainsi que le taux de de change de la Nouvelle-Zélande ont un impact négatif sur le prix du pétrole.

7 Conclusion

L'étude a permis de comparer les différentes techniques de sélection de variables existantes afin de traiter les problèmes de multicolinéarité qui peuvent intervenir lorsque les variables explicatives sont nombreuses, ce qui est le cas pour expliquer les variations des rentabilités du pétrole. Nous avons également comparé ces méthodes aux régressions pénalisées Ridge et Bridge qui ne suppriment pas de variables mais attribuent un poids à chacune d'entre elles selon leur importance. Cependant, les estimateurs peuvent-être moins précis étant donné la quantité de paramètres à estimer. Dans notre cas, il s'est avéré que les régressions pénalisées présentaient un R2 plus élevé que les modèles de sélection avec des taux d'erreurs équivalents. En revanche, pour vérifier la parcimonie nous avons comparé les critères AIC et BIC qui ont montré que cette fois-ci, le modèle GETS était le meilleur modèle puisqu'il minimisait ces critères. Nous avons donc utilisé le modèle le plus parcimonieux pour estimer un modèle par MCO afin de quantifier l'impact des variables selectionnées par GETS sur les rentabilités du prix du pétrole. La vérification de la normalité des résidus, de l'autocorrélation de l'hétéroscédasticité ainsi que de la colinéarité a permis de confirmer la robustesse de notre modèle et affirmer que nos coefficients ne sont pas biaisés.

Table des matières

Table des matières

Table des figures