



# Projet modèle linéaire

*Groupe 8*

JOYMANGUL Jensen

MOISSON Matthieu

20 janvier 2018

Table des matières

[Projet modèle linéaire 0](file:///C:\Users\Matthieu\Downloads\rapport.docx#_Toc504207170)

[I. Presentation du projet 3](#_Toc504207171)

[II. Choix technique d’implémentation 3](#_Toc504207172)

[III. Sélection de variables 4](#_Toc504207173)

[IV. Mesure de performance d’un modèle linéaire 6](#_Toc504207174)

[1. R2 6](#_Toc504207175)

[2. PRESS 6](#_Toc504207176)

[3. Mean absolute error 7](#_Toc504207177)

[4. Critère d'Information d'Akaike (AIC) 7](#_Toc504207178)

[5. Conclusion 7](#_Toc504207179)

[V. Régression « stepwise » 8](#_Toc504207180)

[1. Pseudo-code de la régression stepwise 8](#_Toc504207181)

[6. Résultat de notre modèle 8](#_Toc504207182)

[VI. Régression ‘Least Absolute Deviation’ 11](#_Toc504207183)

[a. Pseudo-code de la régression ‘Least Absolute Deviation’ 11](#_Toc504207184)

[7. Résultat de notre modèle 11](#_Toc504207185)

[VII. Comparaison des résultats 13](#_Toc504207186)

[VIII. Régression simple (meilleur modèle) 15](#_Toc504207187)

[1. Pseudo-code 15](#_Toc504207188)

[2. Comparaison entre chaque mesure de performance 15](#_Toc504207189)

[3. Résultats de notre modèle 16](#_Toc504207190)

[IX. Régression LAD (meilleur modèle) 18](#_Toc504207191)

[a. Pseudo-code 18](#_Toc504207192)

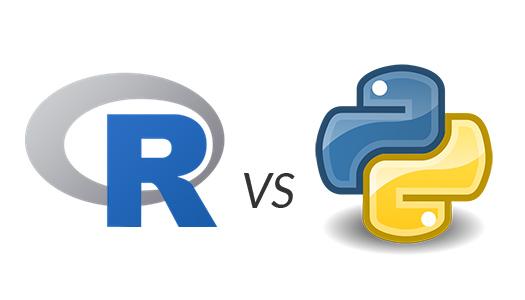
[8. Résultat de notre modèle 18](#_Toc504207193)

[X. Conclusion 20](#_Toc504207194)

1. Presentation du projet

Dans le cadre de la matière de modèle de régression linéaire en Master 2 Data Science à l’université de Lyon 1, nous devions mettre en pratique toutes les techniques ainsi que les théories acquises durant les différents cours magistraux et travaux pratiques. Dans le cadre de la validation de cette matière il nous a été demandé de résoudre un cas réel avec un vrai jeu de données.

Dans ce projet nous avions à notre disposition deux jeux de données distinct sous forme de fichier Excel. Le premier fichier contient 100 observations pour 50 variables. Ce jeu de données sera par la suite utilisé pour la sélection de variable en effet nous allons éliminer les variables redondantes et non pertinentes. Le deuxième fichier lui contient 25 observations avec les 50 même variables que le premier fichier. Dans ce nouveau jeu de données il y a une colonne « réponse » qui contient la réponse obtenue avec les différentes variables. Ainsi, sur ce jeu de données nous allons créer un modèle afin de prédire au mieux la réponse. Pour ce faire nous allons utiliser des méthodes de régression simple.

1. Choix technique d’implémentation

Pour ce projet notre choix concernant l’implémentation s’est porté sur l’utilisation du langage R. Ce choix a été principalement motivé car nous avons déjà utilisé R dans les travaux pratique et cela nous permettra d’approfondir nos connaissances de ce langage. En effet cela nous aidera dans nos stages respectifs.

De plus en explorant les solutions proposées par python, il n’y avait pas de package qui permet de faire la régression « Least Absolute Deviation ». Ainsi si nous avions choisis Python alors nous devions implémentés la régression LAD, cela impliqué un temps plus important dans la réalisation de ce projet pour avoir le même résultat qu’avec le package de R.

1. Sélection de variables

Comme indiqué précédemment nous avons un jeu de données avec 50 variables pour 100 observations. Ainsi le nombre de variables est trop nombreux, nous devons donc essayer d’enlever un certain nombre de variable sans trop perdre d’informations. Nous devons donc trouver les variables qui sont fortement corrélées avec les autres. Pour mettre en évidence les corrélations nous faisons une régression linéaire entre chaque variable. Nous éliminons par la suite les variables qui sont fortement corrélées.

Dans un premier temps nous réalisons des régressions linéaires successive entre deux variables. Nous analysons ensuite le R2 et ce dernier est supérieur à 0.95, nous gardons seulement un des deux variables. En effet si la variable x est fortement corrélée avec y alors nous pouvons éliminer soit x ou y. Pour aller plus loin si x est fortement corrélée avec y qui lui-même est corrélé avec z alors x est fortement corrélé avec z. Nous pouvons donc éliminer deux variables parmi x, y et z. A cette étape nous avons éliminé les 8 variables suivantes :

***descripteur2, descripteur4, descripteur9, descripteur11, descripteur12, descripteur18, descripteur20, descripteur37***

Par conséquence nous nous retrouvons avec 42 variables, ce nombre est encore trop important. Afin d’éliminer plus de variables nous avons encore une fois réalisé des régressions linéaires mais cette fois ci nous faisons une régression linéaire entre 3 variables. Par exemple si nous voulons éliminer la variable y, nous réalisons une régression de y en fonction de deux autre variables (par exemple x et z) et si R2 est supérieur à 0.95 alors nous pouvons conclure que les variables x et z peuvent expliquer la variable y. Par conséquence nous pouvons éliminer la variable y. À cette étape nous avons éliminé les 6 variables suivantes :

***descripteur17, descripteur25, descripteur27, descripteur43, descripteur56, descripteur57***

Dès lors il nous reste 36 variables dans nos jeux de données, nous réitérons le même processus des régressions linéaires avec 4 variables et à la suite de cette étape nous éliminons les 3 variables suivantes :

***descripteur42, descripteur54, descripteur58***

Finalement nous gardons 33 variables. Nous n’avons pas continué d’éliminer d’autres variables car faire des régressions linéaires successifs avec 5 variables, prend beaucoup trop de temps. De plus il y a un risque important de perte d’information. Nous avons DONC pu éliminer 17 variables ce qui représente 34% des variables de départ.

Maintenant que nous avons sélectionner les variables pertinentes, nous allons pouvoir par la suite construire un modèle linaire afin d’expliquer la variable « reponse » grâce au second fichier (observations).

1. Mesure de performance d’un modèle linéaire

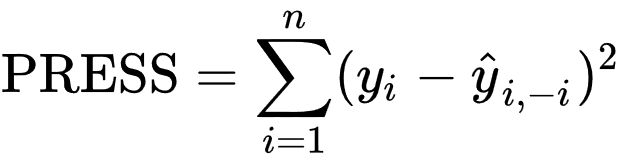
Avant de pouvoir construire notre modèle il nous faut savoir comment évaluer sa performance. Il y a plusieurs manières de mesurer la performance d’une régression la plus simple est celle déjà utiliser dans la sélection de variables qui est l’utilisation du coefficient de détermination (R²). Mais nous allons aussi étudier d’autre marqueur pour étudier la pertinence de notre régression.

1. R2

Le coefficient de détermination (R²) mesure l'adéquation entre un modèle issu d’une régression linéaire simple ou multiple et les données observées qui ont permis de l'établir. Concrètement, ce coefficient se situe entre 0 (le modèle ne vaut rien) et 1 (il est parfait). Par exemple, un coefficient de 0,8 indique que 80 % de la dispersion est expliquée par le modèle de régression. Dans le cadre d’une régression linéaire simple, le plus rapide est d'élever au carré le coefficient de corrélation.

1. PRESS

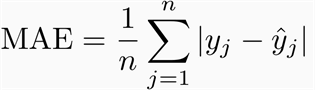
Un autre moyen de mesurer la performance d’une régression est de calculer le « predicted residual error sum of squares (PRESS)». Afin de calculer le PRESS, nous utilisons la formule suivante :



Le PRESS se calcule simplement comme la somme des carrés des résidus(erreurs) de la prévision pour ces observations. Ainsi plus le PRESS est petit plus le modèle est performant. R2 n'offre aucun aperçu significatif sur la façon dont notre modèle de régression peut prédire les valeurs futures. La statistique PRESS peut être utilisée comme une mesure du pouvoir prédictif. Cependant l’inconvénient de cette méthode est que le PRESS est sensible aux grands écarts. En effet si une valeur prédite a une grande écarte par rapport à la réalité, le PRESS augmentera alors même que le modèle a une bonne prédiction pour toutes les autres valeurs.

1. Mean absolute error

Afin de pallier l’inconvénient du PRESS nous pouvons aussi étudier le Mean Absolute Error (MAE). Il est moins sensible à l'erreur occasionnelle très importante car il ne corrige pas les erreurs dans le calcul. MAE mesure l'ampleur moyenne des erreurs dans un ensemble de prédictions, sans tenir compte de leur direction. C'est la moyenne sur l'échantillon d'essai des différences absolues entre la prédiction et l'observation réelle où toutes les différences individuelles ont un poids égal.



1. Critère d'Information d'Akaike (AIC)

Quand un modèle impliquant q paramètres est ajusté aux données, le critère est défini comme suit :

Où Lq est le log vraisemblance maximisé. Akaike a suggéré de maximiser le critère pour choisir entre les modèles dont les nombres de paramètres sont différents.

1. Conclusion

Nous avons fait le choix d’utiliser les trois dernières mesures de qualité du modèle. PRESS, MAE et AIC.

1. Régression « stepwise »

Afin de construire un modèle pour expliquer la réponse, nous avons commencé par utiliser une régression « stepwise ». Le principe de cet algorithme est de faire des régressions successives et à chaque itération nous croisons la meilleure régression et insérons cette variable dans le modèle. Nous répétons cette opération jusqu’à choisir le nombre de variable que nous avons dans le modèle. Dans notre cas nous allons nous arrêter à 3 variables.

1. Pseudo-code de la régression stepwise

n <- 3 //nb de variable dans le modele

list\_modele <- construire la liste de tous les modèles possibles à 3 variable

modele\_courant <- ⃝

tant que n != 0 faire

pour i=0 jusqu`à nb\_variable :

regression\_courant <- lm(reponse~modele\_courant + variable[i])

list\_regression.add(regression\_courant)

fin pour

best\_variable <- choisir meilleur regression dans list\_regression

ajouter best\_variable dans modele\_courant

n<- n - 1

fin tant que

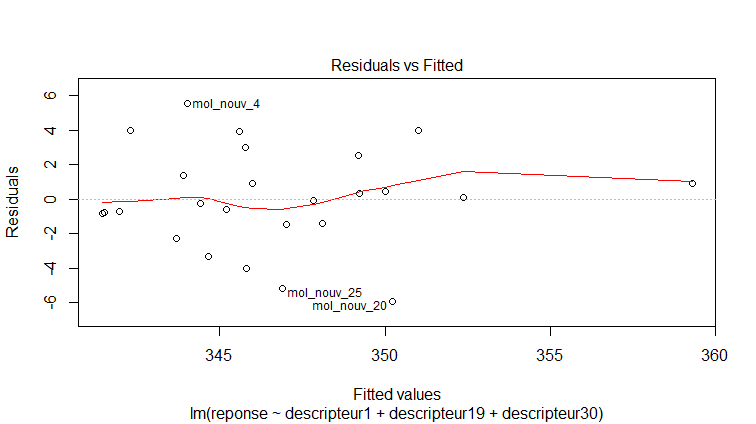
1. Résultat de notre modèle

Après avoir lancer notre régression stepwise, nous avons obtenu le résultat suivant :

|  |  |
| --- | --- |
| Variable 1 | Descripteur1 |
| Variable 2 | Descripteur19 |
| Variable 3 | Descripteur30 |

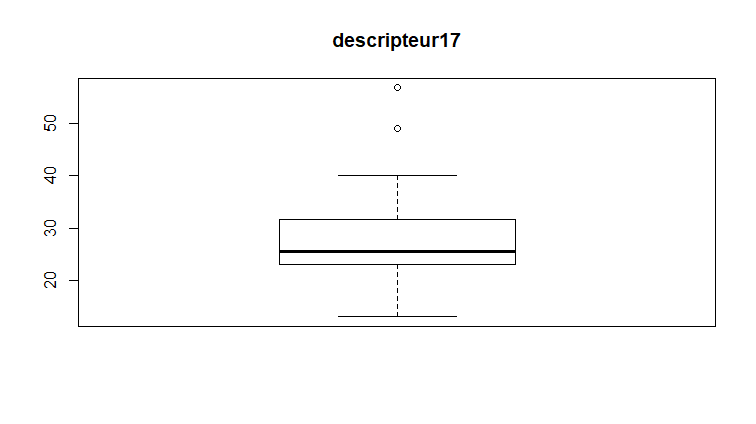
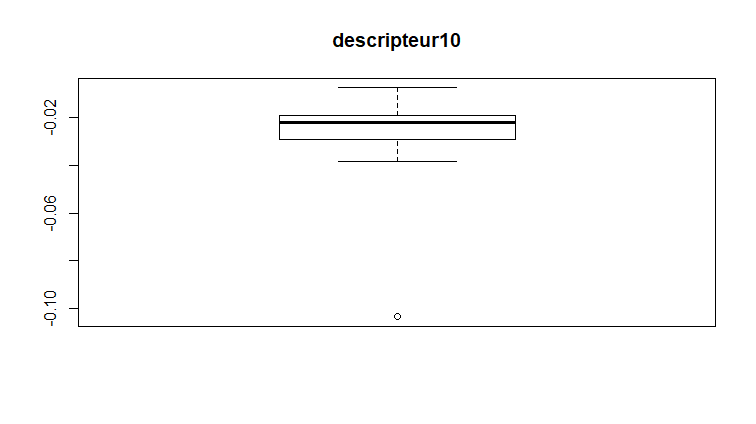
Comme décrit précédemment nous avons calculé le PRESS et le MAE pour le modèle obtenu.

Avec la valeur du PRESS obtenu nous pouvons constater que le modèle obtenu n’est pas très performant. Ce résultat se confirme avec le graph suivant qui représente les résidus contre les valeurs prédits.



Nous constatons que notre modèle n’arrive pas à prédire au plus juste la variable « reponse ». La prédication la plus proche est celle pour l’observation mol\_nouv\_13 ou on a une erreur de seulement 0.054 alors que pour l’observation mol\_nouv\_20 nous obtenons l’erreur la plus importante (erreur = 5.90).

Ce mauvais score peut aussi s’expliquer car dans la majorité des variables il y a des outliers. Dans les boxplot suivant nous pouvons voir les outliers présents dans les descripteurs 10 et 17.



Comme nous avons pu le constater il y a des outliers pour ses deux descripteurs. Nous avons également analyser les autres descripteurs et dans la majorité des descripteurs il y avait des outliers. Cela peut expliquer la performance relativement faible de notre modèle.

1. Régression ‘Least Absolute Deviation’

Afin d’essayer d’améliorer le modèle obtenu avec la régression « stepwise », nous avons implémenter un algorithme pour déterminer un modèle grâce à une régression « Least Absolute Deviation ». Ce type de régression est plus robuste face aux valeur extrême (outliers). Ainsi, cette régression semble plus adaptée dans notre cas.

Le modèle « Least Absolute Deviation » minimise la valeur absolue des résidus :

MAE regression

Ce modèle est très performant même avec des outliers, cependant nous risquons d’avoir plusieurs solutions pour une variable. L’algorithme sélectionnée ressemble à celui du régression stepwise cependant nous calculons des régressions « Least Absolute Deviation ». De plus pour sélectionner le meilleur modèle nous avons utilisé le AIC.

1. Pseudo-code de la régression ‘Least Absolute Deviation’

n <- 3 //nb de variable dans le modele

list\_modele <- construire la liste de tous les modèles possibles à 3 variable

modele\_courant <- ⃝

tant que n != 0 faire

pour i=0 jusqu`à nb\_variable :

regression\_courant <- lad(reponse~modele\_courant + variable[i])

list\_regression.add(regression\_courant)

fin pour

best\_variable <- choisir meilleur regression dans list\_regression

ajouter best\_variable dans modele\_courant

n <- n - 1

fin tant que

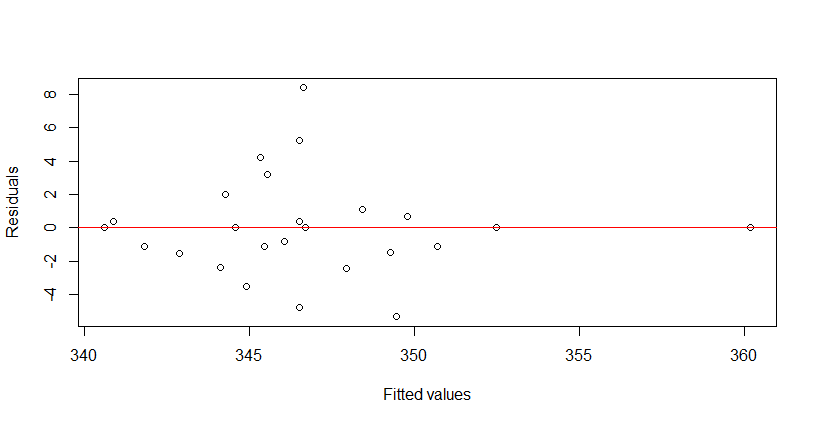
1. Résultat de notre modèle

Après avoir lancer la régression LAD, nous avons obtenu le résultat suivant :

|  |  |
| --- | --- |
| Variable 1 | Descripteur1 |
| Variable 2 | Descripteur19 |
| Variable 3 | Descripteur30 |

Comme décrit précédemment nous avons calculé le PRESS, MAE et AIC pour le modèle obtenu.

Avec la valeur du PRESS obtenu nous pouvons constater que ce modèle est moins performant que le modèle avec la régression stepwise. Nous avons étudié pourquoi le PRESS est plus élevé dans ce modèle par rapport au premier alors que nous avons espéré un meilleur résultat. Nous avons pu constater que le MAE et AIC sont légèrement mieux. Nous reviendrons dessus dans la *section 6 Comparaison des résultats*. Le graph suivant représente les valeurs résiduels(erreur) contre les valeurs prédites par notre modèle.

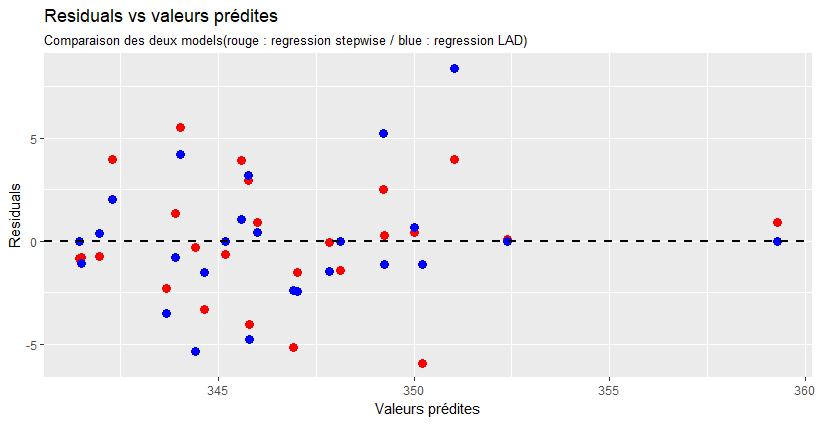


Comme nous pouvons le constater ce modèle arrive à prédire quelques résultats sans aucune erreur, il y a plus exactement 4 prédictions correcte sans erreur. De plus nous pouvons constater que pour certaines prédictions il y a une grande erreur. En effet, pour certaines prédictions il peut il y avoir une différence de 8 entre la valeur réelle et la valeur prédite. Cela peut sembler beaucoup cependant nous pouvons constater que pour la majorité des observations l’erreur est minime. Soit la différence entre la valeur réelle et la valeur prédite est moins de 2.

Ce modèle semble donc être relativement meilleur par rapport au modèle construit par la régression stepwise mais comme indiqué précédemment le PRESS est plus élevé dans ce modèle ci. Il nous faut donc comprendre cette incohérence.

1. Comparaison des résultats

Nous allons maintenant comparer les deux modèles obtenus. Le graphe représente l’erreur (residuals) obtenu pour chacun des deux modèles. Les bulles en rouge représentent le modèle pour la régression stepwise alors que les bulles en bleu représentent le modèle pour la régression « Least Absolute Deviation ».



Comme nous pouvons le constater il y a quelque valeurs prédites ou le modèle avec la régression « Least Absolute Deviation » est moins bon que celui avec la régression « stepwise », Par conséquence ses erreurs influe grandement sur PRESS calculé pour la régression LAD. Donc pour voir la qualité du modèle avec la régression LAD, il faut voir les valeurs du MAE et AIC.

|  |  |  |
| --- | --- | --- |
|  | Régression stepwise | Régression LAD |
| PRESS | 196.783 | 215.1857 |
| MAE | 2.151371 | 2.042 |
| AIC | 132.5276 | 130.359 |

Le MAE et AIC sont légèrement mieux pour la régression LAD, nous pouvons conclure que le modèle avec cette régression est de meilleure qualité cependant ce modèle est loin d’être parfait. De plus avec l’algorithme stepwise nous ne garantissons pas d’avoir le meilleur modèle car c’est un algorithme glouton. Pour aller plus loin nous avons essayer d’implémenter un nouvel algorithme qui permet de calculer le meilleur modèle possible.

1. Régression simple (meilleur modèle)

Les deux algorithmes précédents ne garantissent pas que le modèle linéaire est le plus performant possible. Nous avons donc implémenté deux nouveaux algorithmes afin de garantir que le modèle construit soit de meilleure qualité. Cependant avec ses nouveau algorithmes le temps de calcul est plus long, nous nous retrouvons avec un algorithme de complexité O(n3).

Dans la réalité cette algorithme ne peut être utilise car il est trop lent. Cependant dans notre cas nous nous retrouverons avec un nombre restreint de variables (33) et aussi avec seulement 25 observations donc dans notre cas nous pouvons utiliser cet algorithme. D’ailleurs dans notre cas nous avons :

Nous nous retrouvons avec ce chiffre car nous voulons construire un modèle à 3 variables et si nous voulions construire un modèle avec plus de variable alors le nombre de combinaison possible serai trop grand.

1. Pseudo-code

list\_modele <- construire la liste de tous les modèles possibles à 3 variable

tant que list\_modele pas vide faire

modele\_courant <- list\_modele.pop()

regression\_courant <- lm(reponse~modele\_courant)

si regression\_courant meilleur que regression\_précédent alors

meilleur\_modele <- modele\_courant

fin si

fin tant que

1. Comparaison entre chaque mesure de performance

Pour sélectionner le meilleur modèle, nous testons les trois types de mesures (R2,PRESS et MAE). Le tableau suivant montre les variables obtenus pour chacune des mesures de performance.

|  |  |  |  |
| --- | --- | --- | --- |
|  | R2 | PRESS | MAE |
| Variable 1 | Descripteur1 | Descripteur1 | Descripteur1 |
| Variable 2 | Descripteur23 | Descripteur23 | Descripteur23 |
| Variable 3 | Descripteur61 | Descripteur71 | Descripteur71 |

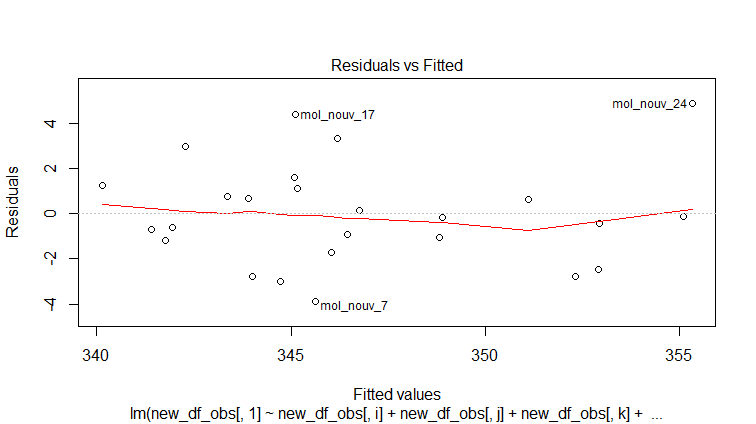
Nous pouvons constater selon la mesure, que le modèle privilégier est différent ainsi la troisième variable diffère. Cependant dans notre cas ce changement n’influe pas dans le résultat final, nous en discutons plus en détails dans la partie suivante. De plus le modèle obtenu avec PRESS et MAE est le même.

1. Résultats de notre modèle

Comme décrit précédemment nous avons obtenus les deux modèles les plus performant selon la mesure utilisée. Cependant quand nous avons analysé plus finement chaque modèle nous avons compris que les trois modèles sont exactement les mêmes. En effet nous avons obtenu strictement les mêmes valeurs de PRESS et MAE pour les trois modèles.

Nous pouvons conclure que les descripteur 61 et descripteur 71 sont fortement corrélés. Nous pouvons donc éliminer l’une des deux descripteurs dans notre jeux de données.

Avec les valeurs du PRESS et MAE obtenus nous pouvons constater que le modèle obtenu est bien plus performant que l’algorithme « stepwise ». Ce résultat se confirme avec le graph suivant qui représente les résidus contre les valeurs prédits.



Nous constatons que notre modèle n’arrive pas à prédire au plus juste la variable « reponse ». Cependant l’erreur est réduite par rapport à l’algorithme « stepwise ». La prédication la plus proche est celle pour l’observation mol\_nouv\_10 ou on a une erreur de seulement 0.09 alors que pour l’observation mol\_nouv\_24 nous obtenons l’erreur la plus importante (erreur = 4.87).

1. Régression LAD (meilleur modèle)

Comme dans l’algorithme précédent nous avons implémenté un algorithme pour trouver le modèle le plus performance possible cependant cette fois-ci nous allons utiliser une régression « Least Absolute Deviation ».

Tout comme dans la l’algorithme précédente, nous ne pouvons pas utiliser cet algorithme dans la réalité car ce dernier est trop lent, nous avons aussi le même nombre de combinaison possible :

1. Pseudo-code

list\_modele <- construire la liste de tous les modèles possibles à 3 variable

tant que list\_modele pas vide faire

modele\_courant <- list\_modele.pop()

regression\_courant <- lad(reponse~modele\_courant)

si regression\_courant meilleur que regression\_précédent alors

meilleur\_modele <- modele\_courant

fin si

fin tant que

Le pseudo-code est quasiment le M6eme que celui de la régression simple (meilleur modèle) cependant ici nous utilisons le régression LAD.

1. Résultat de notre modèle

Après avoir lancer la régression LAD, nous avons obtenu le résultat suivant :

|  |  |
| --- | --- |
| Variable 1 | Descripteur1 |
| Variable 2 | Descripteur8 |
| Variable 3 | Descripteur30 |

Comme décrit précédemment nous avons calculé le PRESS, MAE et AIC pour le modèle obtenu.

Avec ses résultats nous pouvons conclure que ce modèle est moins performant que celui avec la régression simple (meilleur modèle). Ce résultat est surprenant car nous nous attendions que le modèle obtenu avec la régression LAD (meilleur modèle) soit le meilleur résultat. Nous avons émis l’hypothèse qu’avec la régression LAD le modèle est instable et cela peut donner un résultat moins performant dans ce cas.

1. Conclusion

A la fin de ce projet nous avons monté en compétence dans le langage R pour la partie technique. Nous avons compris le fonctionnement de ce langage qui est fortement adapté aux études statistiques. De plus avec cette étude de cas nous avons aussi approfondi nos connaissances théoriques sur la régression et modèle linéaire. Cela nous a permet d’avoir un esprit critique sur les résultats que nous avons obtenus car nous avons compris les résultats que nous obtenions. Cela nous a permis de mettre en pratique les cours théoriques sur un cas réelle.

Concernant le cas en lui-même nous avons obtenu des résultats plutôt satisfaisants par rapport aux jeux de données que nous avions à notre disposition. Malgré que le modèle ne soit pas parfait, il permet de d’expliquer la variable réponse avec beaucoup moins de paramètre et donc cela permet de réduire les coûts dans une possible application réelle. Cependant nous ne pouvons pas prendre le résultat du modèle comme une vérité absolue, nous pouvons faire des simulations pour essayer d’avoir le même résultat pour la variable « reponse » mais cela doit être validé par un teste réelle.

Afin d’améliorer la performance du modèle il y a plusieurs possibilités possibles :

1. Avoir plus de données, c’est-à-dire avoir plus observations
2. Construire un modèle non pas avec 3 variable mais avec un nombre plus grand de variable par exemple un modèle avec 4 ou 5 variable
3. Essayer d’éliminer les outliers avant de construire le modèle (nettoyer les données)