MRR: Rendu 2, Groupe 10

Naïm SOUNI, Abdoulaye SAKHO

Décembre 2021

Introduction

Pour commencer, rappelons que le but de notre projet est d'être capable à partir de différentes données sur un Ouragan à un instant t donnée (en heures), de pouvoir prédire l'intensité de ce même ouragan à l'instant t+24.

De plus, Ce document fait suite au premier rendu de MRR *(Rendu1)* qui décrivait le problème dans sa généralité ainsi que la gestion des données. Nous conseillons vivement au lecteur de lire ce premier document dans un premier temps. Des références au *(Rendu1)* seront faites au cours de ce document.

1 - La modélisation linéaire

Au cours de cette section, on adoptera les notations suivantes: Soit n notre nombre d'observations, et p notre nombre de variables explicatives. Soit Y un vecteur colonne $n * 1$, qui contient les données de notre variable cible. X est notre matrice des données de nos variables explicatives. β est le vecteur des coefficients tel que: $Y = X\beta$.

$\hat{\beta}$ est le vecteur des coefficients calculé par le modèle construit, et $\hat{Y}$ est le vecteur de Y calculé par le modèle construit.

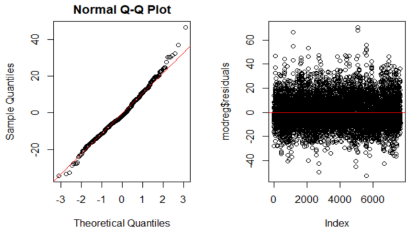
On pose: $E(\beta) = \sum_{i=1}^{n} \epsilon_i^2 = \|Y - X\hat{\beta}\|_2^2$

1.1 Régression linéaire par la méthode des moindres carrés (MCO):

La méthode des MCO a pour but la minimisation de la quantité $E(\beta)$ selon l'argumentβ.En résolvant ce problème d'optimisation, la mérhode MCO nous donne une expression algébrique de $\hat{\beta}$ en fonction de Y et X. Mais ce résultat n'est valable que si la matrice X^TX **est inversble !**

De plus,en supposant que **les résidus sont indépendants et identiquement distribués comme suit:** $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$; on peut retrouver la loi de tous nos estimateurs ($\hat{\beta}$ et$\hat{Y}$). Avec la loi de $\hat{\beta}$, on parvient à construire un test de significativité pour chaque coefficient, c'est le test de Student. Ce test nous permet de savoir si une des variables explicatives du modèle, à travers la valeur de son coefficent donnée, est sinificative ou non(c'est à dire si on peut mettre la valeur de ce coefficient à 0 dans le cas où la variable concernée est non significative). Toutefois, ce test n'a de sens que si **les variables sont peu corrélées. !**.

Dans notre cas, la matrice X^TX est inversible. De plus, la gaussianité et l'homoscédalité des résidus est vérifié. En effet sur la *Figure1*, on observe sur le premier graphique que les quantiles de nos résidus suivent ceux d'une loi normale,l'hypothèse de normalité des résidus est vérifié. Sur le second graphique, on observe une répartion uniforme des résidus de part et d'autre de la droite d'équation y=0, l'hypothese d'homoscédasticité des résidus est verifié. Enfin, nous avons démontré au Rendu2 que nos données compressés sont peu corrélées.



Toutes les hypothèses du modèle linéaire sont vérifié sont vérifié, on peut interpréter les sorties de la fonction *lm()* de **R**.

1.2 Les méthodes incrémentales :

Les méthodes incrémentales sont des méthodes d'optimisation partielle utilisant un critère donnée.

En effet, le principe de ces méthodes est de trouver le meilleur modèle (au sens d'un critère donnée) à partir d'un sous-ensembles de l'ensemble des variables explicatives. Il y aurait en tout 2^p modèles à tester pour trouver le meilleur modèle parcimonieux ! Toutefois, la démarche de ces méthodes permet de ne pas tester toutes les possibilités (c'est pourquoi on les qualifie de modèles d'optimisation partielle).

Un critère est à travers la quantité suivante: $E(\beta, X') = E(MCO(\beta, X') + penalisation(X')$ avec $E(MCO(X')) = \|Y - X'\hat{\beta}\|_2^2$ et X' qui correspond à la matrice des variables explcatives sélectionnées (la pénalisation corresponds au critère).

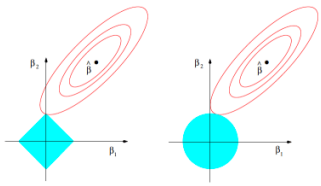
On obtient ensuite: $\beta_{modparcimonieux} = agrmin_{\beta}E(X')$

Les méthodes incrémentales utilisent aussi le test de Student (test de significativité) pour ne pas à avoir à tester toutes les possibilités de modèles parcimonieux. C'est pourquoi ces méthodes supposent les mêmes hypothèse que celle des MCO avec les hypothèses statistiques sur les résidus (résidus gaussien avec homoscédalité, X^TX inversible).

Dans notre cas, on peut donc utiliser ces méthodes. Nous en avons utilisé 3 en tout, le Forward, le Backward et le Bothward avec le critère AIC (ces 3 modèles diffèrent par leur moyen d'approche de la soluton optimale partielle).

1.3 Les méthodes de pénalisation $l1$ et $l2$:

Le principe de ces pénalisation est le même que celui des pénalisation des méthodes incrémentales :
$E(\beta, X) = E(MCO(\beta, X)) + penalisation(X)$.
On obtient ensuite : $\beta_{pen-li} = agrmin_{\beta} E(X)$ et $i \in 1, 2$
Notons qu'ici c'est la matrice X de toutes les variables explicatives qui est utilisée.



Lorsque la pénalisation vaut: $\lambda\|\beta\|_1$; on a une régression Lasso.
Lorsque la pénalisation vaut: $\lambda\|\beta\|_2$; on a une régression Ridge.

Mais contrairement aux méthodes incrémentales, ces méthodes n'utilisent pas le test de Student et ne nécessitent donc pas les hypothèses statistiques sur les résidus du modèle MCO. Toutesfois, ces deux méthodes nécessitent une **renormalisation des données**, illustrée par la *Figure2*. De plus, **l'intercept ne doit pas être pénalisé** ! (Ce que nous avons respecté bien sûr).

Il s'ensuit avec ces deux conditions: $\lambda\|\beta\|_i = \lambda(\sum_{j=1}^{p} \|\beta_j\|_i) \leq 1$ pour $i \in 1, 2$

Ces deux méthodes permettent de gérer les cas dégénérés où la matrice $X^T X$ est non inversible (soit car il existe une dépendance entre les $X^1 ... X^p$ ou soit car on est en grande dimension $p \gg n$). Enfin, notons que la méthode Lasso permet "naturellement" une sélection de variables. En effet, les coefficients du modèle MCO, β_{MCO}, relativement "petits" en norme par rapport à $\lambda/2$ sont mis à 0.

1.4 Les résultats:

Nous avons effectué une 10-Cross Validation particulière sur nos données *(cf Rendu1)*. Chaque type de modèle a été évalué en utilisant cette 10-Cross Validation. On a rassemblé les moyennes des coefficients de détermination et des racines carrées de l'erreur quadratique moyenne de chaque type de modèle dans un tableau (avec la variance de chaque mesure).

RStudio: Notebook Output

Description: df [7 × 4]

	mean(RMSE) <dbl>	var(RMSE) <dbl>	mean(R-squared) <dbl>	var(R-squared) <dbl>
simple	24.83048	7.3622599	0.04231985	1.777095e-02
lm	12.64951	0.4171164	0.74046021	3.198731e-03
Bacward	12.66731	0.4157671	0.73955886	3.155247e-03
Forward	12.66446	0.4197694	0.73920921	1.616284e-05
Both	12.66731	0.4157671	0.73921691	1.594405e-05
Lasso	12.64963	0.4164492	0.73952146	3.201413e-03
Ridge	12.64851	0.4140842	0.73715697	3.162626e-03

7 rows

RStudio: Notebook Output

Description: df [6 × 4]

	mean(RMSE) <dbl>	var(RMSE) <dbl>	mean(R-squared) <dbl>	var(R-squared) <dbl>
lm	12.69944	0.4240635	0.9644352	0.020434849
Bacward	0.00000	0.0000000	0.0000000	0.000000000
Forward	0.00000	0.0000000	0.0000000	0.000000000
Both	0.00000	0.0000000	0.0000000	0.000000000
Lasso	12.45893	0.3537939	0.7499010	0.003697000
Ridge	12.47618	0.3777993	0.7369234	0.003642341

6 rows

La *Figure3* affiche les résultats pour les données compressés *(cf Rendu1)* dans le premier tableau et pour les données non compressées dans le second tableau. La compression a permis d'accélérer grandement le temps d'éxecution des algorithmes et a même permis la convergence des méthodes incrémentales en un temps décent. de plus, les performances des modèles sur les données normalisées sont sensiblement les mêmes que sur les données non normalisées. notre choix de compression des données était donc pertinent.

Notons aussi que les variables sélectionnées par les modèles linéaires de pénalisation restent cohérent avec notre intuition.

1.5 Les limites du modèle linéaire:

Dans un premier temps, on observe sur la *Figure3* que tous nos modèles linéaires (sur les données compressées ou non) restent plus performant que le modèle dit "simple" où on estime que l'itensité de l'ouragan à $t + 24$ est égal à l'intensité de l'ouragan à l'intant t.

Toutesfois, on remarque aussi que tous nos modèles linéaire se trompent en moyenne de 12 noeuds sur la valeur de l'intensité prédite. C'est assez colossale puisque sur l'échelle de Saffir-Simpson, une telle erreur pourait faire passer un ouragan d'une catégorie à une autre. Nous restons donc assez déçu par les modèles linéaires, même en les améliorant avec des pénalisations. Un modèle linéaire n'est peut être pas adapté à la modélisation de notre problème. Cela peut venir d'une relation non évidente et non linéaire entre notre variable cibles et les variiables explicatiives ou bien d'autre chose encore.

2 Pour aller plus loin:

Afin de combler les lacunes d'un modèle dit "strictement linéaire", nous avons décidé d'effectuer un clustering sur nos données et construire un modèle linéaire pour chaque cluster. Les modèles ainsi construit possèdent: **mean(RMSE)** ≈ 9 et **mean(R-squared)** ≈ 0.77.

Nous allons continuer d'essayer de construire des modèles de ce type, c'est à dire non plus "strictement linéaire" car ils dépassent les limites des modèles linéaires comme on vient de le constater.

3 Bibliographie:

The Elements of Statistical Learning: Jerome H. Friedman, Robert Tibshirani et Trevor Hastie.