Projet 1 SY19

Contents

Introduction

Dans le cadre de l'UV SY19, nous avons réalisé un projet consistant en la sélection de modèles de prédictions optimaux pour un problème de régression et un problème de classification. Ce rapport est constitué d'une première partie résumant nos différentes approches pour la régression, puis une deuxième partie concernant la classification.

1 Régression

Avant toute chose, il est utile de comprendre comment sont organisées les données. Nous pouvons remarquer que le jeu de données qui nous est fournit est composé de n = 500 observations pour p = 100 prédicteurs. De plus, la représentation graphique des données en "boxplot" nous permet d'observer que les dispersions et les médianes de chacun des prédicteurs se ressemblent beaucoup. Ces informations nous seront utiles dans le choix des méthodes que nous utiliserons.

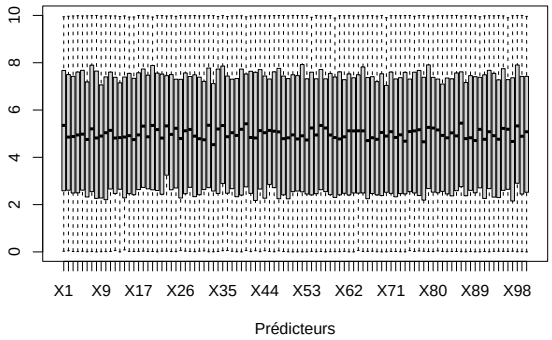

Figure 1: Boxplot des données de régression

Nous pouvons ensuite séparer notre jeu de données en 2 ensembles : un ensemble d'entraînement et un ensemble de test. Ici, nous avont choisi un d'utiliser 4/5 des données totales pour l'ensemble d'entrainement, donc 400 observations d'entraînement pour 100 observations de test, car la quantité d'observations que nous avons à disposition nous le permet.

1.1 Subset Selection avec un critère BIC

Nous avons commencé notre étude du choix d'un modèle par la méthode de "Subset Selection" qui nous permet de trouver un bon sous-ensemble de prédicteurs expliquant les données de régression. Malheureusement, le nombre de prédicteurs étant assez grand, nous n'avons pas pu procéder à une recherche exhaustive des modèles. Il nous fallait alors choisir entre une méthode "forward" ou une méthode "backward". Cette dernière peut en effet fonctionner car n > p. Après plusieurs tests sur notre jeu de données, il s'avère que la méthode "backward" est celle qui retourne l'erreur quadratique la moins grande, c'est donc la méthode que nous utiliserons.

Nous avons ensuite décidé d'utiliser un critère de sélection BIC. Ce dernier nous permet d'ajuster l'erreur d'apprentissage en prenant en compte le biais dû à l'overfitting, tout en pénalisant les modèles avec un grand nombre de prédicteurs. Ce critère nous semble donc utile dans le cas présent car il simplifiera notre modèle.

L'utilisation d'une sélection de sous-ensemble avec un critère BIC nous permet de trouver un modèle performant avec 46 prédicteurs.

1.2 Nested cross-validation sur un modèle linéaire

Nous nous penchons maintenant sur des méthodes de choix de modèle nous permettant directement d'obtenir une estimation de l'erreur de prédiction. Nous mettons alors en oeuvre une validation croisée pour chacun des modèles déterminés avec la méthode de "Subset Selection" présentée dans la partie précédente. Cela devrait nous permettre d'obtenir une bonne estimation de l'erreur de prédiction en fonction du nombre de prédicteurs utilisés. De plus, nous choisissons de découper notre jeu de données d'entrainement en K =

10 sous-ensembles afin d'obtenir un bon compromis biais-variance.

Néanmoins, les résultats obtenus avec cette méthode varient beaucoup en fonction de l'ensemble d'entrainement choisit. Nous procédons donc à une "nested cross-validation" afin de limiter la variabilité de l'erreur. Nous répétons alors 5 fois une validation croisée pour ensuite moyenner les erreurs pour chacun des modèles. On en déduit que le meilleur modèle est un modèle à 51 prédicteurs.

Finalement, nous pouvons aussi utiliser la "One-standard-error rule" afin de choisir un modèle avec moins de prédicteurs, mais qui reste performant. Sur le shéma ci-dessous figure en bleu l'écart-type des erreurs de prédiction ainsi obtenu. Les pointillés verticaux représentent le plus petit modèle respectant cette règle. Il s'agit d'un modèle à 33 prédicteurs.

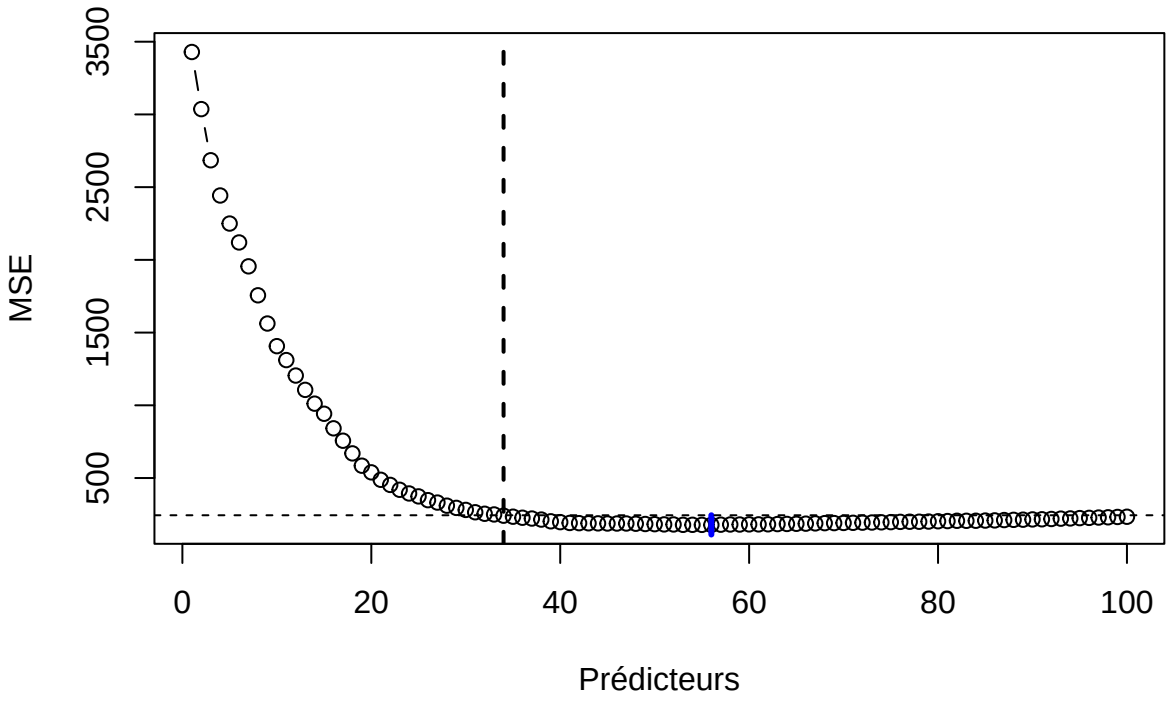

Figure 2: MSE en fonction du nombre de prédicteurs avec une méthode de validation croisée imbriquée

1.3 Régularisation avec une regression Ridge

Nous essayons maintenant d'appliquer les méthodes de régularisation comme la regression ridge pour vérifier si cela ne nous permettrait pas de trouver un meilleur modèle. De plus, cette méthode devrait moins souffrir de la variabilité que la validation croisée.

La variance entre les prédicteurs n'étant pas très grande, il n'est pas nécessaire de centrer et standardiser les données. Nous procédons donc directement à une validation croisée pour choisir la valeur de lambda de sorte à ce que la MSE soit la plus petite possible.

Enfin, nous utilisons la regression Ridge sur les données d'entrainement et calculons l'erreur quadratique moyenne sur les données de test. Les résultats seront présentés dans la partie réservée à la comparaison des modèles.

1.4 Modèle avec méthode lasso

Les résultats obtenus par la regression Ridge comprennent les 100 prédicteurs. Nous utilisons alors maintenant la méthode Lasso afin de trouver un modèle avec moins de prédicteurs. De même, nous utilisons une validation croisée pour trouver une valeur de lambda ayant une MSE faible et un nombre de prédicteurs moins important.

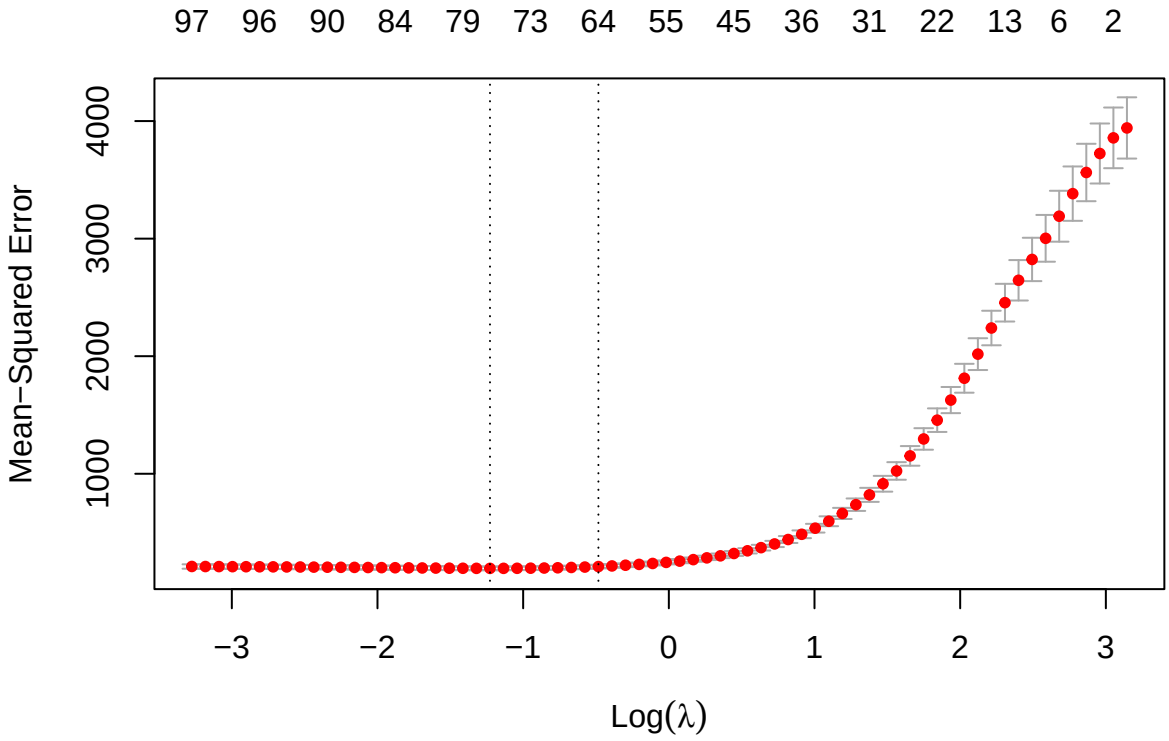

Figure 3: MSE en fonction de la valeur de lambda et du nombre de prédicteurs utilisés

1.5 Comparaison des modèles

Nous cherchons maintenant à determiner quel modèle nous allons utiliser pour répondre à la problématique. Pour cela, nous examinons les résultats de chaque méthode présentés dans le tableau suivant :

Méthode	BIC Subset Selection	Nested Cross-validation	Nested Cross-validation avec la règle "One-standard-error"	Ridge regression	Lasso
MSE sur les données de test	144.18	149,88	207.57	172,83	158.85
Nombre de prédicteurs	46	51	33	100	79

Nous observons alors que deux modèles se démarquent du lot : le modèle trouvé avec le critère BIC qui utilise 46 prédicteurs et le modèle linéaire utilisant 51 prédicteurs trouvé à l'aide de la nested cross-

validation. Ces deux modèles ont presque la même espérance d'erreur quadratique. Le but de ce projet étant de proposer le meilleur modèle possible au sens de la performance, nous choisirons donc le premier pour répondre au problème. En revanche, selon le contexte, il aurait pu être utile de choisir un modèle avec moins de prédicteurs, notamment si la collecte de données est coûteuse. Dans ce cas, nous aurions pu choisir le modèle mis en avant par la méthode de nested cross-validation avec la règle "One-standard-error".

2 Classification