ECP, SDM 2015-2016

Modèles de Régression linéaire
Ridge et Lasso

M. Mougeot, C. Malherbe N. Vayatis

Objectif de la session: Régression avec pénalisation Ridge et Lasso.
Rappel: `help(fonc)` pour obtenir de l'aide sur la fonction nommée "fonc".

Contrôle des connaissances du TD: code R à poster sur Claroline avant le prochain TP.

Application I: Modèle de régression linéaire

- Etudier les fichiers "USCrimeinfo.txt" et "UsCrime.txt". La variable cible (Y) est la première variable colonne du fichier.
- Charger le fichier dans l'environnement R en utilisant la fonction `tab=read.table()`. Quel est le nombre d'observations disponibles? Visualiser les nuages de points entre les variables. Que constate-t-on? Calculer la matrice de corrélations. Interpréter le résultat.

Modèle

On souhaite étudier le modèle linéaire permettant d'expliquer la variable cible en fonction des autres variables disponibles. Expliciter formellement le modèle attendu. La fonction `lm()` de R permet d'estimer les paramètres d'un modèle linéaire et de réaliser différents tests. Consulter l'aide de cette fonction, puis exécuter l'instruction `res=lm('Y~.',data=tab)`, où tab est la structure de type dataframe contenant les données et `res` l'objet contenant le résultat de la fonction. Y correspond au nom de la variable cible étudiée (ici à paramétrer) et `"Y~."` à une formule R spécifiant le modèle étudié.

1. Exécuter les instructions suivantes et noter à chaque fois le résultat proposé: `print(res)`; `summary(res)`; `attributes(res)`;
   En vous aidant des sorties des fonctions précédentes, expliciter le modèle obtenu et les coefficients estimés.

2. **Le modèle globale:** Donner la définition et la signification du coefficient de détermination R^2 ? Quelle est ici sa valeur? Que peut-on en conclure? Ce modèle est-il significatif globalement? Justifier votre réponse à l'aide d'un test statistique approprié.

3. **Les coefficients du modèle:** Tester la significativité de chacun des coefficients du modèle en indiquant le test statistique. Utiliser l'information de p-value. Que peut-on en conclure ici? Les coefficients ont-ils tous le même intéret pour expliquer la variable cible? Justifier votre réponse. Expliquer la signification des codes `***`, `**`, `*` indiqués pour chaque coefficient.

   • Donner un intervalle de confiance pour chacun des coefficients au risque de 5%, puis 1% (`confint()`). Comparer vos résultats à ceux de la question précedente.

4. **Etude des valeurs prédites:** Afficher pour la variable cible les prédictions (en vous aidant des champs de sortie de la fonction lm) en fonction de la valeur cible observée. Que constate-t-on?

   • Calculer les intervalles de confiance pour les valeurs prédites au risque 5%. Deux options sont possibles. Les intervalles de confiance (confidence) correspondent à l'intervalle auquel $E(Y/X = x)$ appartient avec une probabilité de 95%. Les intervalles de prédiction (prediction) correspondent à

l'intervalle auquel, sachant $X = x$, Y appartient avec une probabilité de 95%. Noter que l'intervalle de prédiction prend en compte le terme d'erreur.

5. **Etude des résidus:** Calculer l'erreur quadratique des résidus. Puis, donner une estimation non biaisée de la variance résiduelle.

   • Afficher les résidus $\hat{E}$ (`res$residuals`) en fonction de Y. Que constate-t-on?

   • Etudier la distribution empirique des résidus (`qqnorm`, `qqline`). Le modèle est-il conforme à vos attentes. Justifier votre réponse. • Effectuer un test de normalité des résidus `shapiro.test()`. Conclusion.

6. **Performances du modèle sur de nouvelles données**

   On souhaite à présent évaluer les performances du modèle sur des données non utilisées pour l'estimation des coefficients du modèle. A cet effet, on réalise une partition aléatoire des données.

   - Génerer une liste d'indices, notée indTest, permettant de récupérer successivement une observation sur 3 dans le fichier initial de données en vous aidant de la fonction `seq()`. indTest=1, 3, 6...
   - A l'aide de la variable indTest, réaliser une partition des données initiales en deux dataframe, notée tabTest contenant 1 observation sur 3 et TabTrain contenant le reste des données (soit 2/3 des données)
   - Estimer les paramètres du modèle à l'aide des données TabTrain, puis les prédictions sur les données TabTest en utilisant la fonction `predict()`. Calculer la moyenne des erreurs quadratiques (et son écart-type) sur la base de test. Conclusion.

7. **Analyse graphique**

   - Exécuter les instructions: `x11(); par(mfrow=c(2,2)); plot(res);`.
   - Analyser les graphes proposés.

Sélection de modèles:

Le but est ici de trouver un modèle parcimonieux (utilisant un nombre restreint p_0 de variables $p_0 < p$) tout en proposant un ajustement linéaire acceptable.

1. **Régression Backward.** Exécuter les instructions suivantes:

   ```
   regbackward=step(reg,direction='backward')
   summary(regbackward)
   ```

   Commenter les variables successivement éliminées. Décriver le modèle réduit sélectionné puis comparer le au modèle initial complet.

2. **Régression Forward.** Etudier la fonction `step()` de R. Puis exécuter les instructions suivantes:

   ```
   regforward=step(lm(R~1,data=tab),list(upper=reg),direction='forward');
   summary(regforward);
   ```

   Commenter les variables successivement sélectionnées. Décriver le modèle réduit sélectionné puis comparer le au modèle initial complet, et au modèle sélectionné par la régression backward. Que constatez-vous? Quelles sont les limites de cette approche ?

3. **Régression Stepwise.** Exécuter les instructions suivantes:

```
regboth=step(reg,direction='both')
summary(regboth)
```

Commenter les variables successivement selectionnées puis éliminées. Comparer les trois modèles de sélection.

4. Exécuter l'instruction `formula(s0)` où $s0$ est un objet retourné par la fonction step. Noter que l'instruction `reg0=lm(formula(s0),data=tab);` vous permet automatiquement de réappliquer et d'étudier, `summary(reg0)`, le modèle sélectionné.

5. Etudier l'aide de la fonction step() pour mettre en place une pénalisation de type BIC. Quel est le modèle obtenu? Conclusions.

---

1 Régression Ridge et Lasso

Analyse préliminaire

Les données traitées dans cet exercice sont des indicateurs de développement (économiques, démographiques et sociologiques) aux Etats-Unis sur une période de 15 ans. Parmis ce large panel de données, nous allons réaliser une analyse permettant d'identifer quels indicateurs expliquent, ou du moins sont liés, à la quantité de CO2 émis dans l'atmosphère. Pour ce faire, nous allons d'abord utiliser une régression de type Ridge, juger de la valildité du résultat obtenu ainsi que des limites de cette méthode, puis réaliser une régression de type Lasso afin de réaliser dans un même temps régression et sélection de variables explicatives.

1. Etudier les fichiers "usa_indicators_info.txt" et "usa_indicators.txt".

2. Charger le fichier dans l'environnement R en utilisant la fonction `tab=read.table()`. Comparer le nombre d'observations disponibles au nombre d'indicateurs. Le problème de régression vous semble-t-il simple sur ce type de données ?

3. Quel est la variable contenant la quantité de CO2 émis par an ? Afficher l'évolution de cet indicateur au cours du temps.

4. Les données étant de natures très diverses, les ordres de grandeur peuvent varier notablement d'un indicateur à l'autre. En quoi cela peut-il être un problème pour la régression ? Normaliser les données à l'aide de la fonction `scale(tab, center=FALSE)`.

Régression Ridge

1. Rappeler la définition de la régression de type Ridge.
La fonction `lm.ridge` de la librairie `MASS` de R permet d'effectuer une régression Ridge. Charger la librairie MASS dans l'environnement de travail, puis consulter l'aide de la fonction `lm.ridge`.

2. Effectuer une régression Ridge pour une valeur du paramètre de pénalisation $\lambda = 0$, $\lambda = 100$ (ne pas oublier de retirer la variable `Year` des variables explicatives !). Afficher les coefficients estimés à l'aide de l'instruction `coef()`. Classer par ordre décroissant les coefficients estimés et afficher les 5 plus influents. Ces indicateurs vous semblent-ils vraisemblables ? Retrouver les valeurs des paramètres estimés à l'aide de la régression multiple pour le cas $\lambda = 0$. Noter la différence de résultats entre les instructions `coef(resridge)` et `resridge$coef`.

3. Effectuer une régression ridge pour un ensemble de valeurs de pénalisation λ allant de 0 à 100 par pas de 0.01 ($\lambda = seq(0, 100, 0.01)$).
   Afficher la courbe des performances obtenue par validation croisée en fonction de λ (champ `$GCV` de l'objet résultat, GCV pour Generalized Cross Validation).
   Afficher l'évolution des valeurs des différents coefficients estimés en fonction de λ (`plot(resridge)`). Que constatez-vous ?
   Quel est le modèle à retenir ? Donner la valeur du paramètre de régularisation λ associé. Récupérer automatiquement le modèle recherché à l'aide de la fonction `which.min()`. Récupérer dans la variable `coefridge=...` les coefficients estimés pour ce modèle puis donner leurs valeurs.

4. Calculer l'erreur quadratique moyenne entre les données cibles et les données estimées ($\hat{Y}_{ridge}$), à l'aide d'un produit matriciel où X est la matrice contenant les données des variables explicatives (`Yridge=as.matrix(X)%*%as.vector(coefridge)`).

Régression Lasso

1. Rappeler la définition de la régression de type Lasso. La fonction `lars` de la librairie `lars` de R permet d'effectuer une régression de type Lasso. Charger la librairie dans l'environnement de travail, puis consulter l'aide de la fonction *lars*.

2. Effectuer une régression de type lasso de la variable cible (Y) en fonction des variables explicatives disponibles (`reslasso=lars(X,Y,type="lasso")`. X est de type matrice et contient les variables explicatives, Y est le vecteur de valeurs cibles.
   Exécuter les deux instructions suivantes : `plot(reslasso)` et `plot(reslasso$lambda)`. Que représentent ces graphiques?

3. Récupérer puis afficher la valeur des coefficients pour $\lambda = 0$. Que constatez-vous?
   (`coef=predict.lars(reslasso,X,type="coefficients",mode="lambda",s=0)`)

4. Récupérer, puis afficher la valeur des coefficients pour $\lambda = 0.02$, $\lambda = 0.04$, $\lambda = 0.06$. Que constatez-vous ? Les variables sélectionnées vous semblent-elles plus vraisemblables qu'avec la régression Ridge ?

5. Calculer l'erreur quadratique moyenne entre les données cibles et les données estimées ($\hat{Y}_{lasso}$), et comparer le résutat à l'erreur quadratique de la régression Ridge
   (`pY=predict.lars(reslasso,X,type="fit",mode="lambda",s=0.06)`).