ECP, SDM 2014-2015
M. Mougeot, K. Scaman, N. Vayatis, G. Fäy

Méthodes de Régression

Objectif de la session: Régression linéaire multiple. Sélection de modèle par critère AIC. Régression logistique.
Rappel: `help(fonc)` pour obtenir de l'aide sur la fonction nommée "fonc".

Les fichiers de données ainsi que le support de TD sont disponibles à l'adresse suivante:
`https://sites.google.com/site/mougeotmathilde/teaching`.

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.

Application II: régression logistique

Applications: le fichier "SAHeart.txt" contient les données historiques d'une étude réaliseé sur les facteurs responsables d'un attaque cardiaque pour $n = 462$ habitants d'Afrique du Sud (`http://www-stat.stanford.edu/~tibs/ElemStatLearn`). Le fichier "SAHeartinfo.txt" décrit les variables étudiées et le fichier "SaHeart.txt" contient les données d'étude. Consulter ces deux fichiers. La variable "chd" est la variable de réponse binaire étudiée qui indique si un individu a eu (chd=1) ou pas (chd=0) un incident cardiaque. Les autres variables sont les variables explicatives potentiellement liées à la réponse.

1. Charger les données dans l'environement de travail R. Récupérer dans une structure de type dataframe les variables explicatives suivantes `sbp,tobacco,ldl,famhist,obesity,alcohol,age` et la variable de réponse `chd`. A quoi correspondent ces variables?

2. Visualiser à l'aide d'un scatterplot le jeu de données correspondant aux variables explicatives, en distinguant pour chaque individu le type de réponse (1/0) à l'aide d'un code couleur.

   (`pairs(tab,pch=22,bg=c("red","blue")[unclass(factor(tab[,"chd"]))])`).

3. **Régression logistique:** La fonction `glm()` de R permet d'estimer les paramètres d'un modèle linéaire généralisé. Consulter l'aide cette fonction. Utiliser cette fonction pour estimer les paramètres du modèle. Utiliser la fonction `summary()` pour une description complète de l'objet R.

4. Comparer pour l'ensemble des individus la réponse prédite et la réponse attendue. Calculer la matrice de confusion et le pourcentage de "faux positifs" $P(\hat{Y} = 1/Y = 0)$ et de "faux négatifs" $P(\hat{Y} = 0/Y = 1)$. Conclusion. (fonction `table()`)

5. **Validation croisée:** On souhaite à présent estimer les coefficients du modèle sur 75% des individus (base d'apprentissage), puis évaluer la qualité des résultats obtenus par ce premier modèle sur les 25% des individus restants (base de test). Calculer la matrice de confusion sur les bases de test et d'apprentissage. Répeter cette procédure plusieurs fois et estimer l'erreur min, max, moyenne de classification. Quel est l'intéret d'une telle approche?

6. Effectuer une régression logistique avec sélection de variables en utilisant la fonction `step`. Quels sont les coefficients retenus les plus significatifs, les moins? Que peut-on en conclure?

   Calculer les performances de ce nouveau modèle? conclusion.