UNIVERSITÉ DE BRETAGNE-SUD

Master MSAD : Apprentissage statistique et Big-Data

Examen du 25 mai 2022

Durée : 2 h

On considère les données suivantes :

Y	1	2	4	5	3.5	3	2.1	1.5	0.8
x	0.2	0.7	0.8	1.2	1.4	1.6	2.4	2.5	2.9

On cherche f telle que $Y = f(x)$.

Régression linéaire simple

1. Représenter Y en fonction de x et effectuer une régression linéaire. Ecrire le modèle (on identifiera la matrice X) et commenter soigneusement les résultats.
2. On note $\widehat{Y}$, les réponses fournies par ce modèle.
   Montrer formellement que le biais est nul en moyenne
   i.e. $\sum_{i=1}^{n}(\widehat{Y} - Y_i) = 0$. Calculer avec R.
3. Donner l'expression de la matrice de variance-covariance de $\widehat{Y}$ en fonction de σ^2. Calculer les coefficients de σ^2 sur la diagonale de la matrice.
4. Calculer la variance empirique de $\widehat{Y}$.
5. Rappeler l'expression et calculer une estimation de l'erreur quadratique moyenne (MSE).
   Que faudrait-il faire pour avoir une meilleure estimation de cette erreur ?
6. Appliquer la technique LOOCV *(Leave-One-Out Cross-Validation)* pour approximer le MSE.

Régression non linéaire

On se propose de rechercher un modèle non linéaire de la forme :

$$Y_i = \beta_1 f_1(x_i) + \beta_2 f_2(x_i) + \beta_3 f_3(x_i) + \varepsilon,\ \varepsilon \sim \mathcal{N}(0,\sigma^2),\ i = 1,\ldots,n. \qquad (1)$$

Différents choix pour les fonctions $f_j(x)$, $j = 1,2,3$, sont possibles et il s'agit d'estimer $\beta = (\beta_1, \beta_2, \beta_3)$.

1. On considère la partition de l'intervalle $[0,3]$ en trois intervalles A_j tels que $A_j = [j-1, j[$, $j = 1,2,3$. et des fonctions f_j de la forme $f_j(x) = 1$, si $x \in A_j$ et $f_j(x) = 0$, sinon, $j = 1,2,3$.
   (a) Ecrire le modèle (1) pour les f_i définies sous forme matricielle (On écrira soigneusement la matrice X).
   (b) Rappeler la formule des estimateurs de moindres carrés donnant le vecteur $\widehat{\beta}$.
   (c) Calculer $\widehat{\beta}$ à partir de la forme matricielle. Faire les calculs avec R.
   (d) Représenter le modèle (fonction en escalier).
   (e) Montrer de manière formelle, que le biais est nul.
   (f) Exprimer la variance en fonction de σ^2 et comparer avec le résultat obtenu à la question 3 de la partie précédente. Calculer son estimation empirique avec R.
   (g) Appliquer la technique LOOCV pour approximer le MSE.
   (h) Comparer ce modèle avec le précédent.
2. On se propose d'ajuster des modèles linéaires dans chacun des intervalles A_j, $j = 1,2,3$.
   (a) Effectuer les régressions pour les valeurs dans chacun des intervalles A_j, $j = 1,2,3$. Commenter.
   (b) Représenter la régression ainsi obtenue. Commentaires
   (c) Calculer la variance empirique des prédicats et appliquer la technique LOOCV pour approximer le MSE.
3. Effectuer une régression sur un polynôme de degré 2.
   (a) Commenter les résultats et représenter le modèle.
   (b) Calculer le MSE par la technique LOOCV.
   (c) Comparer avec une régression polynomiale de degré 3.

Régression splines

La régression splines est une technique qui permet d'ajuster une modèle non linéaire en considérant des polynôme pour les $f_i(x)$ et en imposant une contrainte de continuité. Le code suivant effectue une régression splines sur les données de l'énoncé :

```
library(splines)
fit=lm(Y~bs(x,knots=c(1,2)))
pred<-predict(fit,newdata=list(x=seq(0.2,2.9,0.01)),se=T)
plot(x,Y, col="blue", pch=20)
lines(seq(0.2,2.9,0.01), pred$fit,col="red")
```

Appliquer ce code et commenter le résultat en comparant avec ce qui a été obtenu précédemment. On représentera le modèle.