Apprentissage de données biologiques 2025-2026
Contrôle des connaissances (durée : 1h)

21 octobre 2025

Nom Prénom :

Tous les documents sont autorisés. La calculatrice est autorisée. L'usage du téléphone portable est interdit, même pour sa fonction calculatrice.

L'évaluation de la teneur en Carbone Organique d'un sol est un enjeu important, pour la quantification des niveaux de pollution. La méthode la plus fiable pour mesurer cette teneur en Carbone Organique repose sur une analyse physico-chimique coûteuse, qui n'est pas envisageable pour une évaluation rapide du niveau de pollution d'une surface importante. D'autres méthodes sont donc envisagées, dont la spectrométrie proche infra-rouge.

Afin d'évaluer la précision de cette méthode, on procède à une expérience sur une parcelle située dans le sud tunisien, sur un chott. Sur cette zone, on définit une grille régulière de 144 points de mesure sur lesquels on prélève un échantillon de sol pour une analyse déterminant sa teneur en CO (g/kg). Chaque échantillon de sol est également soumis à spectrométrie proche infra-rouge (on dispose d'une mesure spectrale par nm entre 400 et 2500 nm).

On importe les données dans un objet appelé `dta`, dont la première colonne, nommée `OC`, contient les teneurs en carbone organique. Le tableau suivant donne un aperçu de la répartition de la teneur en carbone organique et des valeurs spectrales (après transformation SNV) pour les 5 premières longueurs d'ondes :

```
summary(dta[,1:6])
```

```
       OC           nirs400          nirs401          nirs402      
 Min.   :0.123   Min.   :-3.77   Min.   :-3.77   Min.   :-3.81  
 1st Qu.:0.456   1st Qu.:-3.65   1st Qu.:-3.64   1st Qu.:-3.65  
 Median :0.676   Median :-3.59   Median :-3.59   Median :-3.60  
 Mean   :0.704   Mean   :-3.52   Mean   :-3.52   Mean   :-3.52  
 3rd Qu.:0.916   3rd Qu.:-3.50   3rd Qu.:-3.49   3rd Qu.:-3.49  
 Max.   :2.387   Max.   :-1.82   Max.   :-1.83   Max.   :-1.83  
    nirs403          nirs404      
 Min.   :-3.80   Min.   :-3.78  
 1st Qu.:-3.66   1st Qu.:-3.64  
 Median :-3.59   Median :-3.59  
 Mean   :-3.52   Mean   :-3.51  
 3rd Qu.:-3.48   3rd Qu.:-3.48  
 Max.   :-1.82   Max.   :-1.79  
```

L'objectif est donc de construire une règle de prédiction de la teneur en carbone organique à partir des spectres proches infra-rouges.

Question 1

Dans cette problématique, quelle est la variable à expliquer et quelles sont les variables explicatives ? Vous donnerez la nature (quantitative ou catégorielle) de ces variables et le nombre de variables explicatives.

Réponse

Question 2

Proposez un modèle statistique permettant d'étudier la relation entre la teneur en carbone organique et l'analyse par spectrométrie proche infra-rouge d'un échantillon de sol. Votre réponse doit contenir le nom du modèle et l'équation mathématique reliant la variable à expliquer aux variables explicatives.

Réponse

Dans une première approche, on cherche à construire un modèle qui permettrait de prédire la teneur en carbone organique à partir d'une partie des variables explicatives, choisie judicieusement mais limitée à un maximum de 50 variables.

Pour cela, on utilise la fonction `regsubsets` du package `leaps` de `R` :

```
fullmod <- lm(OC~.,data=dta)
select <- leaps::regsubsets(OC~.,data=dta,nvmax=50,
                method="forward")
select <- summary(select)
optimal_number <- which.min(select$bic)
select <- lm(OC~.,data=dta[,select$which[optimal_number,]])
```

Question 3

Des deux modèles `fullmod` et `select` estimés ci-dessus, pour lequel la somme des carrés des résidus est-elle la plus petite ? Selon quel critère le modèle `select` est-il meilleur que le modèle `fullmod` ?

Réponse

Les commandes suivantes donnent les résultats de deux procédures de validation croisée pour évaluer la précision de la prédiction par la méthode de sélection implémentée ci-dessus.

```
n <- nrow(dta)
seg <- pls::cvsegments(n,k=10)
cvpred.select <- rep(0,times=n)

for (k in 1:10) {
   mod <- lm(formula(select),data=dta[-seg[[k]],])
   cvpred.select[seg[[k]]] <-
     predict(mod,newdata=dta[seg[[k]],])
}
select1 <- sqrt(mean((dta$OC-cvpred.select)^2))
print(select1)
```

```
[1] 0.0730241
```

```
for (k in 1:10) {
   fullmod <- lm(OC~.,data=dta[-seg[[k]],])
   select <- leaps::regsubsets(OC~.,data=dta[-seg[[k]],],
                               nvmax=50,method="forward")
   select <- summary(select)
   optimal_number <- which.min(select$bic)
   select <- lm(OC~.,data=dta[-seg[[k]],select$which[optimal_number,]])
   cvpred.select[seg[[k]]] <-
     predict(select,newdata=dta[seg[[k]],])
}
select2 <- sqrt(mean((dta$OC-cvpred.select)^2))
print(select2)
```

```
[1] 0.266031
```

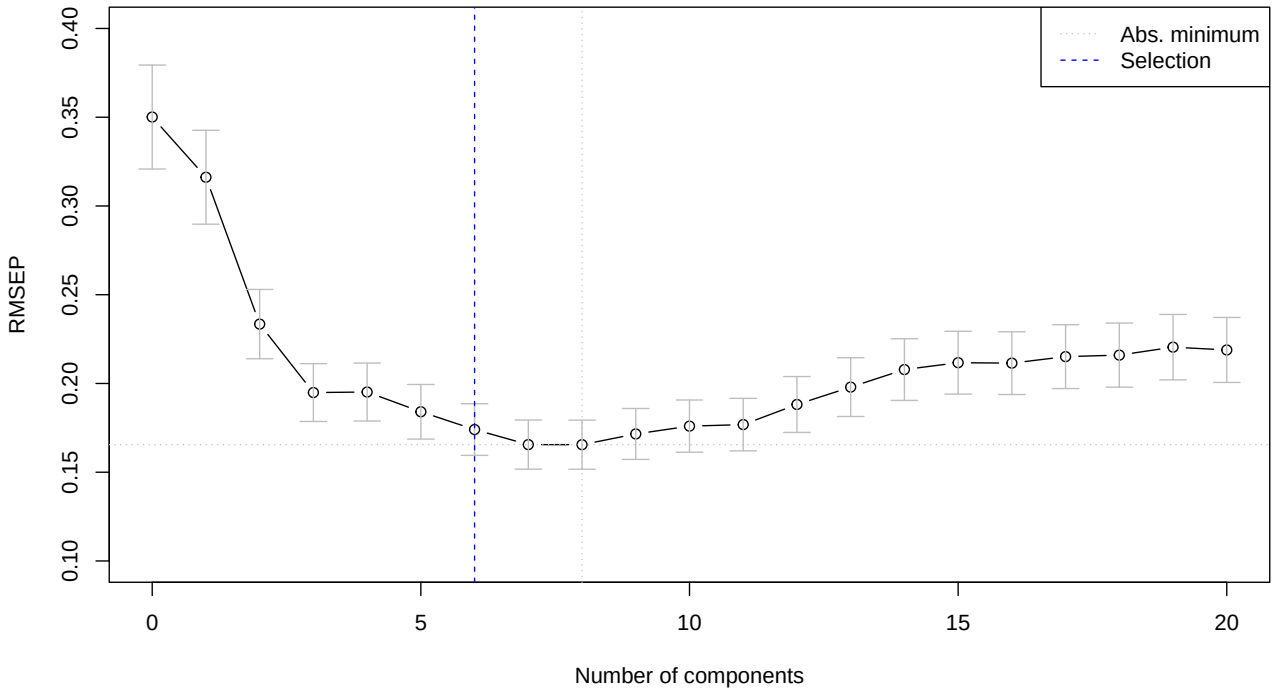
Question 4

Que mesurent les quantités `select1` et `select2` ? Pourquoi la valeur de `select2` est-elle plus élevée que celle de `select1` ?

Réponse

Une des méthodes d'estimation possibles pour estimer les paramètres du modèle de la question 2, la régression `Partial Least Squares (PLS)`, s'appuie sur l'extraction d'un certain nombre q de variables latentes à partir des données de spectres. Les commandes suivantes donnent les résultats de l'ajustement du modèle par la méthode PLS implémentée dans le package `pls` de `R`.

```
OC.cvpls <- pls::plsr(OC~.,data=dta,
                      ncomp=20,
                      validation="CV",segments=n)
nb_comp <- pls::selectNcomp(OC.cvpls,plot=TRUE,ylim=c(0.1,0.4),
                            method="onesigma")
```




```
OC.pls <- pls::plsr(OC~.,data=dta,ncomp=nb_comp)
```

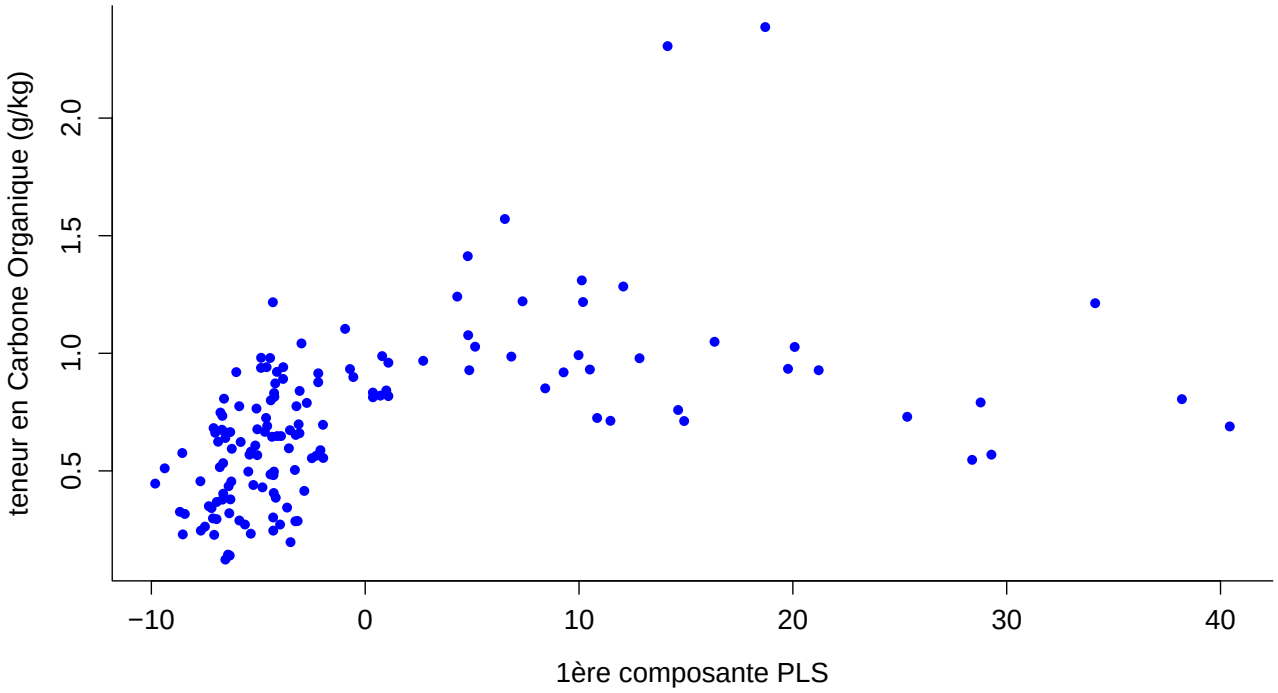
Question 5

Quelle est la méthode de validation croisée utilisée ci-dessus pour estimer les écarts-types d'erreurs de prédiction (RMSEP) associés à chaque choix d'un nombre de composantes PLS ? Diriez-vous que la PLS avec un choix optimal de nombre de composantes conduit à une précision satisfaisante de la prédiction ? (Justifiez votre réponse).

Réponse

Le graphique suivant permet de visualiser le lien entre la teneur en Carbone Organique et la 1ère variable latente.

```
LV <- predict(OC.pls,type="scores") # Prédit les composantes PLS
colnames(LV) = paste("lv",1:nb_comp,sep="")
plot(LV[,1],dta$OC,type="p",pch=16,col="blue",
     xlab="1ère composante PLS",
     ylab="teneur en Carbone Organique (g/kg)",bty="l",
     cex.lab=1.25,cex.main=1.25,cex.axis=1.25)
```



Question 6

Pourquoi la figure ci-dessus remet-elle en question une hypothèse de la méthode PLS ?

Réponse

Les commandes suivantes montrent les résultats de l'ajustement d'un modèle de prédiction de la teneur en Carbone Organique, s'inspirant de la méthode PLS mais en corrigeant le problème révélé dans la question précédente.

```
oc.gam <- gam::gam(OC~s(lv1,df=8)+lv2+lv3+lv4+lv5+lv6,
    data=data.frame(OC=dta$OC,LV))
```

Question 7

Donnez le nom et l'équation mathématique du modèle ajusté ci-dessus, reliant la teneur en carbone organique aux variables latentes. En quoi diffère-t'il du modèle de la méthode PLS standard ?

Réponse

Les commmandes suivantes donnent les résultats d'une procédure de validation croisée permettant d'évaluer la précision de la prédiction du modèle discuté ci-dessus.

```
cvpred <- rep(0,n)
for (i in 1:n) {
   OC.pls <- pls::plsr(OC~.,data=dta[-i,],ncomp=nb_comp)
   LV <- predict(OC.pls,type="scores") # Prédit les scores
   colnames(LV) = paste("lv",1:nb_comp,sep="")
   OC.gam <- gam(OC~s(lv1,df=8)+lv2+lv3+lv4+lv5+lv6,
       data=data.frame(OC=dta$OC[-i],LV))
   LV_i <- predict(OC.pls,newdata=dta[i,],type="scores")
      # Prédit les composantes PLS pour l'individu i
   colnames(LV_i) = paste("lv",1:nb_comp,sep="")
   cvpred[i] <- predict(OC.gam,newdata=
            data.frame(OC=dta$OC[i],LV_i))
}
rmsep <- sqrt(mean((dta$OC-cvpred)^2))
rmsep
```

```
[1] 0.175683
```

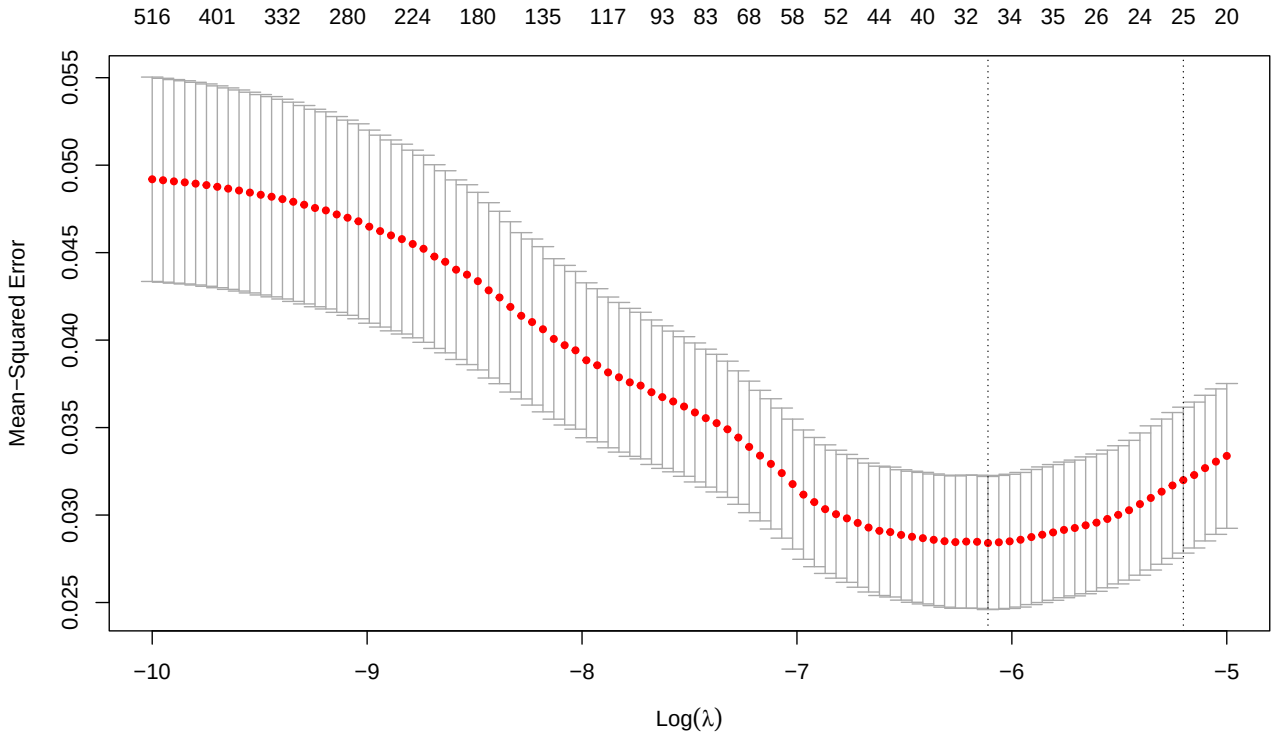
Question 8

Diriez-vous que la modification discutée ci-dessus de la méthode de régression PLS a conduit à une amélioration de la performance de prédiction ? (Justifiez votre réponse).

Réponse

On cherche maintenant à voir s'il est possible d'atteindre voir d'améliorer ce niveau de précision dans la modélisation de la teneur en carbone organique à partir d'une autre approche de type 'régression pénalisée', la méthode LASSO. La Figure ci-dessous montre l'évolution de l'erreur quadratique moyenne de prédiction en fonction du paramètre de pénalisation.

```
log_lambda <- seq(from=-10,to=-5,length=100)
OC.cvglmnet <- glmnet::cv.glmnet(y=dta$OC,x=as.matrix(dta[,-1]),
   type.measure="mse",lambda=exp(log_lambda))
plot(OC.cvglmnet)
```




```
print(sqrt(min(OC.cvglmnet$cvm)))
```

```
[1] 0.168526
```

Question 9

Le graphique ci-dessus montre que la prédiction de la teneur en carbone organique est la moins précise pour de faibles valeurs du paramètre de pénalisation. Ce défaut de précision est-il dû à une trop forte variance ou à un trop fort biais de l'erreur de prédiction ? (justifiez votre réponse).

Réponse

Question 10

D'après vous, faut-il choisir la méthode LASSO dans le cas présent ? (justifiez votre réponse).

Réponse