Statistique Numérique et Analyse de Données
Ecole des Ponts ParisTech, 2ème année

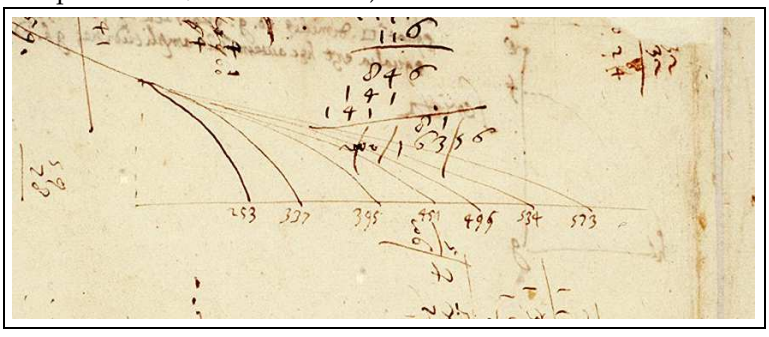
TP3 : RÉGRESSION LINÉAIRE MULTIPLE : ANALYSE DE VARIANCE ET SÉLECTION DE MODÈLE

Sélection de modèle

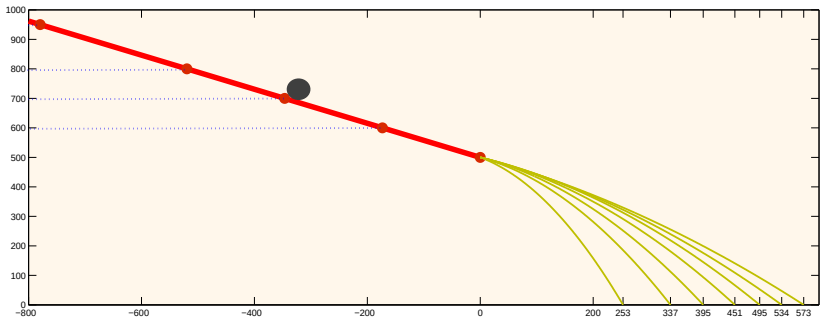
Rappelons que dans une régression linéaire multiple on cherche à prédire/expliquer une variable réponse à l'aide de p variables explicatives. Le but de sélection de modèle est de réduire au maximum l'ensemble des variables explicatives tout en préservant la qualité prédictive/explicative du modèle.
Il y a au moins deux raisons pour s'intéresser à la sélection de modèle. La première découle du célèbre principe philosophique appelé rasoir d'Occam : *Les multiples ne doivent pas être utilisés sans nécessité*. Aussi appelé « principe de simplicité », le rasoir d'Occam est également formulé de la façon suivante : *si deux théories expliquent également bien un résultat, il convient de « trancher » en faveur de la plus simple*.
La deuxième raison motivant l'intérêt à la sélection de modèle est plus quantitative : les modèles contenant un grand nombre de variables explicatives ont tendance à conduire vers un surapprentissage (overfitting). Autrement dit, le modèle basé sur un grand nombre de variables se focalise sur l'explication des valeurs observées au détriment de l'explication du phénomène général.
L'exercice suivant a pour but d'illustrer le phénomène de surapprentissage et ses inconvénients.

Exercice 1. Expérience de Galileo Galilei

Au debut du 17 ème siècle, Galilei a effectué un certain nombre d'expériences visant à étudier les lois physiques décrivant le mouvement d'un corps dans des différentes conditions. Une de ces expériences consistait à mesurer la distance horizontale parcourue par un objet placé à différentes hauteurs sur un plan incliné, ce dernier se trouvant à une hauteur de 500 punti du sol (un punto correspond à 169/180 millimètre).



Le manuscript de Galilei, disponible sur le site
`http://www.mpiwg-berlin.mpg.de/Galilleo_Prototype/HTML/F114_V/C114_V.HTM`



1. Commencer par charger et afficher les données :

```
library(UsingR)
data(galileo)
g2=galileo$init.h-500
g1=galileo$h.d
par(bg='cornsilk')
plot(g1, g2, pch=20, col="red", cex=2, ylim=c(0,1000))
```

   ▶ Au vu du nuage des points obtenu, est-il raisonnable de chercher une relation affine entre les variables `h.d` et `init.h` ?

2. On s'intéresse au problème suivant : ayant mesuré la distance parcourue par l'objet avant de toucher le sol déterminer la hauteur initiale à laquelle l'objet a été laché. Autrement dit, on cherche à déterminer `init.h` en fonction de `h.d` .

   Pour illustrer l'importance de la sélection de modèle, supposons que seuls les 4 premières observations sont disponibles. Nous allons examiner deux façons de déterminer une fonction f telle que `init.h` $\approx f($`h.d`$)$ et, ensuite, nous comparerons la qualité prédictive de ces deux fonctions sur les 3 observations restantes.

   (a) On cherche d'abord un ajustement quadratique : $f(x) = ax^2 + bx + c$. Pour cela, on effectue une régression de `h.d` sur le vecteur `(init.h, init.h^2)`.

```
f1=g1[1:4]
f2=g2[1:4]
LinReg1=lm(f2 ~ f1+I(f1^2))
summary(LinReg1)
```

   ▶ Quelles sont les valeurs estimées des paramètres a, b et c ?
   ▶ Que vaut le coefficient de détermination ?

   (b) On cherche ensuite un ajustement par une fonction f qui s'écrit comme $f(x) = ax^2 + bx + c + d\exp(x/20)$.

```
LinReg2=lm(f2 ~ f1+I(f1^2)+I(exp(f1/20)))
summary(LinReg2)
```

   ▶ Quelles sont les valeurs estimées des paramètres a, b, c et d ?
   ▶ Que vaut le coefficient de détermination ?

   On constatera que le coefficient de détermination est meilleur que pour l'ajustement quadratique. Cela implique-t-il que la deuxième fonction trouvée est meilleure que la première ?

(c) Pour mieux répondre à la question précédente, on calcule et affiche les prédictions fournies par chacune des deux fonctions :

```
t=(2500:6000)/10
new=data.frame(f1=t)

pred1=predict(LinReg1, new, interval ="none")
pred2=predict(LinReg2, new, interval ="none")

par(bg='cornsilk')
plot(g1,g2,pch = 20,col="black",cex=2,ylim=c(0,1000))
points(t,pred1,pch = 20,col="blue",cex=0.2)
points(t,pred2,pch = 20,col="red",cex=0.2)
```

- Quelle est votre conclusion ?

(d) L'exemple précédent met en évidence l'importance du choix de modèle. Il montre également que, pour choisir le modèle, on ne peut pas raisonner composante par composante. En d'autres termes, la stratégie éliminant du modèle toutes les variables explicatives déclarées inutiles par le test de Student n'est pas acceptable. Pourquoi ?

3. Nous allons maintenant utiliser le package `leaps` de `R` pour effectuer la tâche de sélection de modèle. Cela se fait de la manière suivante :

```
f1=g1[1:5]
f2=g2[1:5]
explicative=matrix(c(f1,f1^2,exp(f1/20)),ncol=3)
leaps(x=explicative,y=f2)
```

- Que fait l'instruction `explicative=matrix(c(f1,f1^2,exp(f1/20)),ncol=3)` ?
- Au vu des valeurs de la quantité `Cp` pour chaque modèle, lequel des 7 modèles possibles vous auriez choisi ?

Remarque 1. L'exemple de cet exercice est choisi pour des fins pédagogiques uniquement. En pratique, il faut éviter de mener une analyse statistique sur des données aussi peu nombreuses.

Exercice 2. Critères de sélection : Cp, AIC, BIC.

On s'intéresse aux données `geophones` qui contiennent les deux variables suivantes :

`distance` – la distance du géophone [1] d'un point de référence. Les différents emplacements de l'appareils sont allignés.

`thickness` – l'épaisseur du substratum [2] aux l'emplacements où les expériences ont été réalisées.

On cherche à déterminer un modèle expliquant l'épaisseur en fonction de la distance. Pour fixer les idées, on cherchera à déterminer une fonction polynomiale de degrée 9.

1. Commençons par lire les données :

1. Appareil utilisé en géologie pour déterminer l'épaisseur d'un substratum en mesurant le temps que met un signal pour le traverser
2. En géologie, un substratum désigne une couche inférieure sur laquelle repose une couche plus récente.

```
library(DAAG)
data(geophones)
d=geophones$distance
t=geophones$thickness
par(bg='cornsilk',pch=19,col='red')
plot(geophones)
```

et par former la matrice des variables explicatives :

```
explicative=matrix(c(d,d^2,d^3,d^4,d^5,d^6,d^7,d^8,d^9),ncol=9)
```

- ▶ Quelle est la valeur du coefficient de détermination lorsqu'on effectue une régression de la variable `t` sur `explicative` ?
- ▶ Quel modèle obtient-on si l'on retire du modèle complet toutes les variables explicatives qui sont déclarées inutiles par le test de Student ?

2. On se propose maintenant d'effectuer une sélection de modèle par le biais de la commande `leaps` en minimisant le critère `Cp`.
   - ▶ Quel est le meilleur modèle selon ce critère ?
3. On veut se convaincre que les autres critères conduisent vers des résultats similaires. Pour cela, sésir

```
AIC(lm(t~explicative[,1:9]))
AIC(lm(t~explicative[,2:9]))
AIC(lm(t~explicative[,5:9]))
AIC(lm(t~explicative[,3:9]))
AIC(lm(t~explicative[,1:7]))
```

   - ▶ Quel est le meilleur modèle parmi les 5 modèles ci-dessus selon le critère AIC ?
4. Pour le critère BIC, il suffit d'utiliser la commande `AIC` avec l'option `k=log(nombre de variables)` :

```
AIC(lm(t~explicative[,1:9]), k=log(9))
AIC(lm(t~explicative[,2:9]), k=log(8))
AIC(lm(t~explicative[,5:9]), k=log(5))
AIC(lm(t~explicative[,3:9]), k=log(7))
AIC(lm(t~explicative[,1:7]), k=log(7))
```

   - ▶ Quel est le meilleur modèle parmi les 5 modèles ci-dessus selon le critère BIC ? Ce résultat, est-il différent de ceux obtenus dans les deux questions précédentes ?

Exercice 3. Analyse de variance à un facteur

Nous allons effectuer une analyse de variance à un facteur sur les données de qualité d'air de New York contenues dans `airquality`.

1. Commençons par charger les données :

```
data(airquality)
help(airquality)
D=airquality
```

```
summary(D)
par(bg='cornsilk',col='red')
plot(D,pch=19)
```

Nous voulons d'abord déterminer si les variations de la concentration d'ozone d'un mois à l'autre ont été significatives ou pas. Pour cela, on effectue une ANOVA à un facteur en considérant la variable `Month` comme variable qualitative :

```
LR = lm(Ozone ~ as.factor(Month), data=D)
anova(LR)
```

- Quel est le role de l'instruction `as.factor` ?
- Peut-on affirmer au seuil de 5% que la concentration d'ozone n'a pas varié au fil des mois ?

2. Nous voulons maintenant savoir si le jour du mois a une influence sur la concentration d'ozone ou pas.
   - Ecrire les instructions nécessaires pour répondre à la question ci-dessus et déterminer si oui ou non (au seuil de 5%) la concentration d'ozone dépend de la date ?