TD 1

L'objectif de ce TD est de vous familiariser avec les objets et les commandes du logiciel R relatives au modèle linéaire. Dans la suite, sont écrites à gauche des commandes à taper, et à droite des commentaires sur ces commandes ou des questions relatives. Vous pouvez donc ainsi apprendre les commandes du logiciel par la pratique. Ouvrir un « script » dans lequel vous pourrez sauver toutes les commandes passées et les exécuter.

Le symbole > apparaît automatiquement en début de chaque ligne de commandes.

Le symbole + apparaît en début de ligne si la précédente est incomplète.

Le symbole # permet d'insérer un commentaire.

Une petite astuce très utile lorsque vous tapez des commandes directement dans la console : en utilisant les flèches Haut et Bas du clavier, vous pouvez naviguer dans l'historique des commandes tapées précédemment, que vous pouvez alors facilement réexécuter ou modifier.

Les commandes élémentaires à connaître

`help()`	
`help.start()`	L'aide au format "html" (web)
`q()`	Quitter R
`example(plot)`	
`demo()`	

Premiers pas

`x=c(1,4,9)`	La fonction $c()$ concatène des scalaires ou des vecteurs.
`y=c(x,2,3)`	
`c(1:4)`	
`seq(10,100,10)`	Le premier terme est 10, le dernier est ≤ 100 et le pas est 10.
`x=rep(0,10)`	On crée un vecteur constitué de 0 répété 10 fois.
`rep(y,10)`	
`x=rnorm(20)`	On simule 20 v.a. i.i.d. suivant la loi normale standard.
`y=rexp(20)`	On simule 20 v.a. i.i.d. suivant la loi exponentielle d'espérance 1.
`median(x)`	
`mean(x)`	
`var(x)`	
`sd(x)`	
`summary(x)`	
`sum(x)`	
`length(x)`	
`plot(x)`	Pour tracer la première "courbe"
`lines(x)`	Pour ajouter une ligne
`points(y)`	Pour ajouter un nuage de points
`hist(x)`	
`boxplot(x)`	
`barplot(x)`	

Exercice 1
Tracer la courbe $f(x) = sin(x)$ pour $x \in [-2\pi, 2\pi]$.

Exercice 2
Pour une loi normale d'espérance 5 et de variance 2 :
1. Faire une représentation graphique de sa fonction de répartition et de sa densité sur $[0, 10]$.
2. Calculer la probabilité des évènements : $X \leq 0$, $X \leq 5$, $-1 < X \leq 3$ et $X > 10$.
3. Calculer entre quelles valeurs 95% des tirages de X sont compris.
4. Même question que la précédente, mais pour une loi normale centrée réduite.

Exercice 3
Soient $(x_1, y_1), \cdots, (x_n, y_n)$ des couples de réels. Déterminer le réel $\widehat{a}$ qui minimise la somme des carrés résiduelle $\mathrm{SCR}(a) = \sum_{i=1}^{n} |y_i - a \cdot x_i|^2$ (Indication : on peut dériver la fonction $a \mapsto \mathrm{SCR}(a)$). Que représente $\widehat{a}$ par rapport à $y_1, \cdots, y_n$? Comparer avec la valeur $\widehat{\beta}$ obtenue dans le cadre de la régression linéaire simple classique.

Codes R du cours

```
age=c(35,45,55,65,75)
```
On crée un vecteur de données appelé âge ;

```
tension=c(114,124,143,158,166)
reg=lm(tension~age)
```
Régression de la variable tension par la variable age,

```
summary(reg)
```
Résultats de la régression ;

```
plot(age,tension,type="p")
```
Graphique ;

```
abline(reg,col="red")
```
Trace la droite de régression ;

```
x = as.data.frame(cbind(tension,age))
```
On crée une table de données ;

```
p1=predict(reg,x,interval="confidence",level=0.8,
se.fit=TRUE)
```
Intervalle de confiance ;

```
p2=predict(reg,x,interval="prediction",level=0.8,
se.fit=TRUE)
```
Intervalle de prédiction ;

```
x11()
```
Ouvrir une nouvelle fenêtre pour afficher la figure suivante ;

```
plot(age,tension, xlab='age', ylab= 'tension')
abline(reg,col="red")
points( p1$fit[,2] ~ age, type='l', lty="dotted")
points( p1$fit[,3] ~ age, type='l', lty="dotted")
points( p2$fit[,2] ~ age, type='l', lty="dashed" )
points( p2$fit[,3] ~ age, type='l', lty="dashed" )
legend("topleft", c("Bande de confiance",
"Bande de prédiction"),
lwd=1, lty=c("dotted", "dashed"),cex=0.8)
plot(reg,which=1)
```
Plot des résidus ;

```
plot(reg,which=2)
```
QQ-plot des résidus ;

TD 2
Exemple de régression simple

Etude de simulation sur les intervalles de confiance

On travaille toujours sur l'example de `Tension Age`. On va simuler 100 jeux de données selon le modèle estimé. Les données simulées sont considérées comme les observations des différentes expériences. La méthode des moindres carrés est appliquée à chaque jeu pour obtenir les paramètre $\hat{\mu}$ et $\hat{\beta}$. On obtiendra ainsi 100 lignes droites de régression. On compare finalement les intervalles de confiance obtenues à partir du jeu de données original et les résultats de simulation.

```
# Extraction des paramètres à partir du jeu de données original
mu=reg$coef[1]
beta=reg$coef[2]
sig=summary(reg)[[6]]

# Simuler un jeu de données
y=mu+beta*age+rnorm(5,0,sig)
reg2=lm(y~age)
summary(reg2)
plot(age,tension,type="p") # Tracer le nuage de points des données originales
points(age,y,col="green") # Ajouter le nuage de points des données simulées
abline(reg,col="red") # Tracer la droite de régression originale
abline(reg2,col="cyan") # Tracer la droite de régression avec les données simulées

# Simuler 100 jeux de données et tracer les droites de régression
for (i in 1:100){
      y=mu+beta*age+rnorm(5,0,sig)
      reg2=lm(y~age)
      points(age,y,col="green")
      abline(reg2,col="cyan")
}

# Retracer le modèle original
points(age,tension,type="p");
abline(reg,col="red",lwd=2)
```

Exercice 1

Rajouter une nouvelle donnée (par exemple à 25 ans) dans la table précédente. Suivant la valeur de la tension, regarder les résultats.

Exercice 2

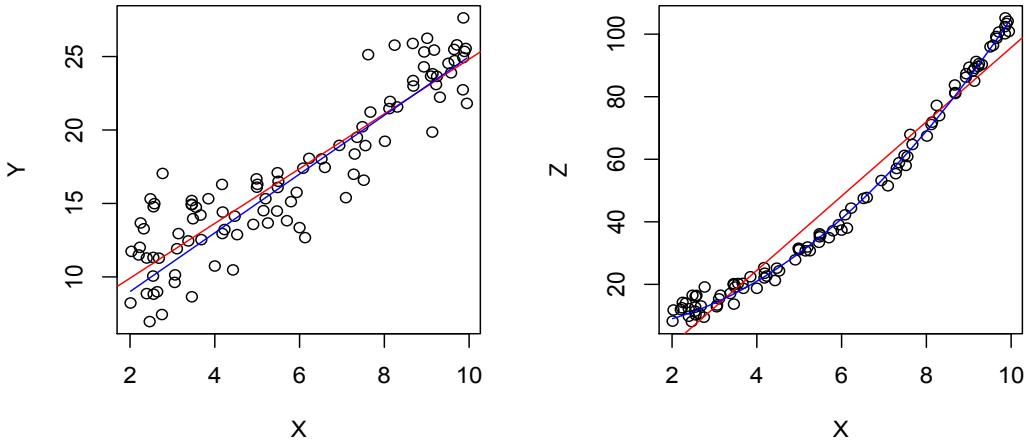
On considère deux jeux de données (X_i, Y_i) et (X_i, Z_i), $i = 1, \ldots, n$, définis par les modèles suivants,

$$Y_i = 2X_i + 5 + \mathcal{N}(0, 4),$$

$$Y_i = X_i^2 + 5 + \mathcal{N}(0, 4).$$

Comparer par simulation les résultats de régression linéaire. Que peut on conclure sur les graphiques des résidus ? Appliquer la régression linéaire sur les données transformées en logarithme. Qu'obtient on ?

1. Générer un vecteur `X` de 100 nombres aléatoires distribués uniformément dans $[2, 10]$.
2. Générer un vecteur `e` de 100 nombres aléatoires selon une loi normale d'espérance 0 et de variance 4.
3. Définir les vecteur `Y=2*X+5+e` et `Z=X*X+5+e`.
4. Estimer les paramètres des droites de régression en utilisant `lm`. Afficher les résultats de la régression.
5. Tracer côte à côte les nuages de points (X_i, Y_i) et (X_i, Z_i), $i = 1, \ldots, 100$. Ajouter les droites de régression en rouge et les lignes de modèle réel en bleu.
6. Afficher les graphiques des résidus de deux modèles dans les fenêtres séparées.
7. Refaire les procédures précédentes, mais sur les couples $(\log X_i, \log Y_i)$ et $(\log X_i, \log Z_i)$.



TD 3
Calcul du pseudo - R^2

On modifie une observation de tension dans l'exemple de régression simple du TD 2, et refait la régression.

```
age=c(35,45,55,65,75)
# On modifie le vecteur de tension
tension=c(114,124,143,156,166)
Tens=data.frame(age,tension)
reg=lm(tension~age,data=Tens)
summary(reg)
av1=anova(reg)
av1
par(mfrow=c(1,2))
plot(age,tension)
abline(reg)
```

On travaille cette fois ci sur les données logarithmiques.

```
lnage=log(age)
lntens=log(tension)
lnTens=data.frame(lnage,lntens)
reg=lm(lntens~lnage,data=lnTens)
summary(reg)
av2=anova(reg)
av2
plot(lnage,lntens)
abline(reg)
```

Le modèle log-linéaire est il meilleur que le modèle simple ?

```
# Calcul du pseudo-R^2
lny=reg$fitted.values
y=exp(lny)
e=tension-y
scr=sum(e^2)
pR2=1-scr/sum(av1$Sum)
pR2
```

Conclusion ?

Exercice 1

Comment calculer le coefficient de détermination à partir de la table d'analyse de la variance ? Vérifier R^2 de deux modèles en utilisant les sorties de l'appel `anova(reg)`.

Exercice 2 : Exemple « Oeufs »

Une entreprise de production d'oeufs produit 10000 oeufs par jour. Compte tenu des contrat qu'elle a passés avec ses clients, il lui faut calibrer cette production d'oeufs en fonction de leur poids. Ce que fait notre entreprise, qui dispose d'une chaîne d'emballage automatisée incorporant une unité de pesage et de tri. Mais cette chaîne est extrêmement lente. La société recherche un procédé économique de calcul, même indirect et approximatif, du poids de ses oeufs, qui permette le pesage-tirage en continu.

Un procédé pourrait être la détermination de la hauteur ou de la largeur de l'oeuf, par des moyens faisant appel à la technique des faisceaux laser. Le poids de chaque oeuf serait déduit de la hauteur ou de la largeur.

Le fichier de données "oeufs.txt" est téléchargeable à l'adresse suivante.

http ://samm.univ-paris1.fr/IMG/txt/oeufs.txt

Enregistrez le dans votre répertoire de travail.

`getwd()`	Récupère le répertoire courant
`setwd("/chemin/Donnees")`	Spécifie le répertoire courant

Le poids de l'oeuf se déduit-il approximativement de sa hauteur ou de sa largeur ? Mieux de l'un ou de l'autre ? Et comment ?

```
oeufs=read.table("/chemin/Donnees/oeufs.txt",header=TRUE)
reg.oeufs=lm(POIDS~HAUTEUR,oeufs)
summary(reg.oeufs)
anova(reg.oeufs)
plot(oeufs$HAUTEUR,oeufs$POIDS)
abline(reg.oeufs)

lnh=log(oeufs$HAUTEUR)
lnp=log(oeufs$POIDS)
lnOeufs=data.frame(lnh,lnp)

reg.lnOeufs=lm(lnp~lnh,data=lnOeufs)
summary(reg.lnOeufs)
anova(reg.lnOeufs)
plot(lnh,lnp)
abline(reg.lnOeufs)

lny=reg.lnOeufs$fitted.values
y=exp(lny)
e=oeufs$POIDS-y
scr=sum(e^2)
pR2=1-scr/(7498.7+282.2)
pR2
```

TD 4
Régression linéaire multiple

Colinéarité statistique

Nous vérifions d'abord la formule $\hat{\theta} = (X'X)^{-1}X'Y$ pour le modèle simple.

```
X=cbind(rep(1,5),age);X
Y=matrix(tension,nrow=5,ncol=1);Y
solve(t(X)%*%X)%*%(t(X)%*%Y)
```

Nous estimons les paramètres pour un modèle multiple en utilisant la formule.

```
data(AirPassengers);AP = AirPassengers
temps = time(AP);t = (temps-1955)^2

# Création des variables muettes
vm = matrix(0,nr = length(AP), nc = 12)

for(i in 1:12) {
v = rep(0,12)
v[i] = 1
vm[,i] = rep(v,12)
}

mod = lm(log(AP)~temps+t+vm[,1]+vm[,2]+vm[,3]+vm[,4]+vm[,5]+vm[,6]
+vm[,7]+vm[,8]+vm[,9]+vm[,10]+vm[,11]+vm[,12])
summary(mod)

# Vérification
t1 = matrix(temps,nrow=144,ncol=1);t2 = matrix(t,nrow=144,ncol=1)
X = cbind(rep(1,144),t1,t2,vm);X
Y = matrix(log(AP),nrow=144,ncol=1);Y
solve(t(X)%*%X)%*%(t(X)%*%Y)

# Suppression de la dernière variable
X = cbind(rep(1,144),t1,t2,vm[,1:11]);X
solve(t(X)%*%X)%*%(t(X)%*%Y)
```

Exercice 1

On travaille sur les données "oeufs.txt". En utilisant la formule $\hat{\theta} = (X'X)^{-1}X'Y$ pour estimer le modèle `POIDS ~ HAUTEUR` et `POIDS ~ HAUTEUR+LARGEUR`.

Nous allons montrer les paradoxes possibles de la significativité partielle et le problème de la colinéarité statistique par l'exemple ci dessous. Nous avons deux séries d'observations sur des variables Y, X_1 et X_2. Le nombre d'observations est le même dans les deux cas ($n = 10$).

```
# colinéarité statistique
Y=c(22.73,21.36,24.09,26.89,24.95,32.88,33.54,31.39,36.85,30.69)
X1=c(10,11,12,13,14,15,16,17,18,19)
X2=c(14,12,11,14,11,18,18,15,15,12)

y=c(21.99,21.37,24.72,27.16,30.60,31.52,33.35,38.21,33.55,40.29)
x1=c(10,11,12,13,14,15,16,17,18,19)
x2=c(12,11,12,13,16,15,16,17,18,21)

reg1=lm(Y~X1+X2)
reg2=lm(y~x1+x2)
summary(reg1)
summary(reg2)

cor(X1,X2)
cor(x1,x2)
```

TD 5
Maximum de vraisemblance et test de Student

Exercice 1 Maximum de vraisemblance

On rappelle que, dans le cadre d'un modèle statistique paramétrique (c'est-à-dire que la loi du modèle ne dépend que d'un nombre fini de paramètres), la densité des observations $Y_1, \ldots, Y_n$ vue comme une fonction des paramètres est appelée la vraisemblance. L'estimateur du maximum de vraisemblance de ces paramètres est la valeur des paramètre qui maximise la vraisemblance.

a) Considérons le modèle gaussien

$$Y_i = \mu + \beta X_i + \varepsilon_i, \quad i = 1, \ldots, n$$

et ses paramètres (μ, β, σ^2). Montrer que $-2 \times$ log-vraisemblance vaut

$$L(\mu, \beta, z) = n \log(2\pi) + n \log z + \frac{1}{z} \sum_{i=1}^{n} (Y_i - \mu - \beta X_i)^2,$$

en notant $z = \sigma^2$ pour dériver plus facilement.

b) Comparer les estimateurs du maximum de vraisemblance avec ceux des moindres carrés de μ et β.

c) Comparer (au sens de la vitesse de convergence) l'estimateur du maximum de vraisemblance de σ^2 avec $\hat{\sigma}^2$ défini dans le cours.

Exercice 2 Test du rapport de vraisemblance

On rappelle que dans le cadre de variables aléatoires $Y_1, \ldots, Y_n$ suivant un modèle statistique admettant une densité, la statistique U du rapport de vraisemblance est le rapport entre le maximum (en les paramètres) de la vraisemblance sous l'hypothèse H_0 et le maximum (en les paramètres) de la vraisemblance sous l'hypothèse H_1. La région d'acceptation du test est du type $U \leq \lambda$, avec $\lambda \in \mathbb{R}$ dépendant du niveau du test.

a) On considère le test du rapport de vraisemblance de $H_0 : \beta = 0$ contre $H_1 : \beta \neq 0$ dans le cadre du modèle gaussien de l'exercice 1. Montrer que si l'on pose

$$SC := \sum_{i=1}^{n} (Y_i - \hat{\mu} - \hat{\beta} X_i)^2,$$

alors le maximum sous H_1 de la log-vraisemblance vaut

$$-\frac{1}{2}(n \log(2\pi) + n \log(\frac{SC}{n}) + n).$$

b) Montrer un résultat équivalent sous H_0 en posant

$$\tilde{SC} := \sum_{i=1}^{n} (Y_i - \bar{Y})^2.$$

c) Donner l'expression de $\log(U)$ et comparer avec les tests de Student et Fisher.

Exercice 3 Tests sur la moyenne d'un échantillon gaussien

Un vigneron veut savoir quelle est la contenance moyenne des bouteilles qu'il produit. Il effectue pour cela une mesure sur un échantillon de 20 bouteilles et obtient, en centilitres, les volumes suivants.

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
76	75	77	74	73	77	72	74	74	73	76	74	77	73	75	75	78	72	72	77

Il déduit que la moyenne empirique est 74.7. Supposons que la contenance d'une bouteille prise au hasard dans le production est distribuée selon une loi normale de moyenne μ et de variance σ^2. Alors peut on croire que $\mu = 75$?

σ connu

```
# X_1,...,X_20 ~ N(mu,2)
x=c(76, 75, 77, 74, 73, 77, 72, 74, 74, 73, 76, 74, 77, 73, 75, 75, 78, 72, 72, 77)
n=length(x)
# l'écart-type de la statistique est 2/sqrt(n)
q=qnorm(0.975,0,1); w1=75-q*2/sqrt(n); w2=75+q*2/sqrt(n);
mean(x)<w1 | mean(x)>w2
st=(mean(x)-75)*sqrt(n)/2; p=2*(1-pnorm(abs(st)))
```

σ inconnu

```
# dans ce cas mean(x) centré réduit suit une loi de student à n-1 degré de liberté
sd=sqrt(var(x)); q=qt(0.975,n-1); w1=75-q*sd/sqrt(n); w2=75+q*sd/sqrt(n);
mean(x)<w1 | mean(x)>w2
t=(mean(x)-75)*sqrt(n)/sd; p=2*(1-pt(abs(t),n-1))

# fonction t.test()
t.test(x); t.test(x, mu=75); t.test(x,mu=75,conf.level=0.5)
```

Comment calculer l'intervalle de confiance ? La moyenne μ est elle supérieure à 74 ?

Exercice 4 Étude de cas : BACTER

L'évolution au cours du temps du nombre Y de bactéries dans des conditions données de température de milieu a été observée heure par heure en laboratoire. Les résultats obtenus sont les suivants.

T	0	1	2	3	4	5	6	7	8	9	10
Y	2	3	4	5	6	8	10	13	16	22	28

Les chercheurs ayant réalisé ces mesures pensent que deux hypothèse sont vraisemblables.

H1 : La variation du nombre de bactéries est proportionnelle au temps écoulé.

H2 : Le taux de variation du nombre de bactéries est proportionnel au temps écoulé.

Question 1 : Quel modèle se « cache » derrière chacune de ces deux hypothèses ?

Question 2 : Jugez par tous les moyens possibles de la validité de ces hypothèses. Quelle hypothèse faut il considérer comme la plus vraisemblable ?

Question 3 : La pente β est elle supérieure à 1 dans le premier modèle et supérieure à 0.2 dans le second modèle ?

Question 4 : Quel pronostic faites vous pour la 25ème heure ?

TD 6

Exercice 1

Soit Y qui suit un modèle linéaire non-gaussien, et soit $T \in \mathbb{R}^n$ un vecteur déterministe. Montrer que

$$\mathbb{E}\left(\|T - Y\|^2\right) = n\sigma^2 + \|T - X\theta\|^2.$$

Exercice 2 Théorème de Gauss-Markov

On se propose de montrer dans cet exercice que, pour le modèle linéaire non gaussien, l'estimateur des moindre carrés $\hat{\theta}$ reste optimal mais maintenant seulement parmi les estimateurs linéaires sans biais.

L'optimalité veut dire que si $\tilde{\theta}$ est un autre estimateur linéaire sans biais :

$$\mathrm{Var}\left(\tilde{\theta}\right) - \mathrm{Var}\left(\hat{\theta}\right) \text{ est une matrice semi-définie positive,}$$

ou encore, ce qui est équivalent, que pour toute combinaison linéaire $C'\theta$ des paramètres où C est un vecteur de même taille que θ, i.e., $C \in \mathbb{R}^{k+1}$

$$\mathrm{Var}\left(C'\tilde{\theta}\right) \geq \mathrm{Var}\left(C'\hat{\theta}\right).$$

a) Posons $\tilde{\theta} = MY$ où M est une matrice de taille $(k+1, n)$. Montrer que $MX = I$.

b) Écrire $\hat{\theta} = TP_{[X]}Y$, et montrer que $MP_{[X]} = TP_{[X]}$.

c) Montrer que $\tilde{\theta} = \hat{\theta} + MP_{[X]^\perp}Y$, les deux termes de la somme étant non-corrélés. Conclure.

Exercice 3

Soit θ_1 et θ_2 deux paramètres réels inconnus et soit :

— Y_1 un estimateur sans biais de $\theta_1 + \theta_2$ et de variance σ^2;

— Y_2 un estimateur sans biais de $2\theta_1 - \theta_2$ et de variance $4\sigma^2$;

— Y_3 un estimateur sans biais de $6\theta_1 + 3\theta_2$ et de variance $9\sigma^2$,

les estimateurs Y_1, Y_2 et Y_3 étant indépendants. Quels estimateurs de θ_1 et θ_2 proposeriez vous ? (on pourra utiliser l'exercice précédent).

Exercice 4 Modèle passant par l'origine

Considérons le modèle gaussien

$$Y_i = \beta X_i + \varepsilon_i, \quad i = 1, \ldots, n,$$

où $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$, i.i.d. et X_i sont déterministes. Rappelons que la vraisemblance est la densité des observations $Y_1, \ldots, Y_n$ vue comme une fonction des paramètres β et σ^2. La densité d'une variable aléatoire gaussienne $\mathcal{N}(\mu, \sigma^2)$ est

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}.$$

1. Montrer que $-2 \times$ log-vraisemblance vaut

$$L(\beta, z) = n \log(2\pi) + n \log z + \frac{1}{z} \sum_{i=1}^{n} (Y_i - \beta X_i)^2,$$

en notant $z = \sigma^2$ pour dériver plus facilement.

2. Montrer que les estimateurs du maximum de vraisemblance de β et σ^2 sont

$$\hat{\beta} = \frac{\sum_{i=1}^{n} X_i Y_i}{\sum_{i=1}^{n} X_i^2}, \quad \widehat{\sigma^2} = \frac{1}{n} \sum_{i=1}^{n} (Y_i - \beta X_i)^2.$$

3. Les estimateurs $\hat{\beta}$ et $\widehat{\sigma^2}$ sont ils sans biais ? Calculez les variances de $\hat{\beta}$ et $\widehat{\sigma^2}$. (Rappel : $\mathbb{E}(\chi^2(n)) = n$, $\mathrm{Var}(\chi^2(n)) = 2n$.)

4. On dispose de deux jeux de données suivants.

X	8	11	5	6	4
Y	17	21	10	11	8

X	47	43	58	45	36
Y	95	87	118	89	74

Lequel choisissez vous pour estimer β ? Au moment du recueil de données, que devons nous faire pour diminuer les variances de $\hat{\beta}$ et $\widehat{\sigma^2}$?

Exercice 5 Bande de prédiction

Que fait ce programme ? Pouvez vous indiquer la valeur de `n` ?

```
# création des données
e=rnorm(20,mean=0,sd=0.5); x=runif(20,min=-5,max=5); y=2*x+1+e

# estimation du modèle
model=lm(y~x); summary(model)

# prédiction
newdata=data.frame(x=10); predict(model,newdata)

# intervalles de prédiction
p.int=predict(model,newdata,interval="prediction")
int1=p.int[2]; int2=p.int[3]

# simulation de 1000 réalisations de y_10
ys=2*10+1+rnorm(1000,mean=0,sd=0.5); n=0
for (i in 1:1000) if ((ys[i]<int1) | (ys[i]>int2)) n=n+1
n
```

TD 7
Sélection de variables

Exercice 1 Un exemple pédagogique
Considérons les sept variables suivantes qui pourrons être des variables explicatives.

Variables	$X^{(1)}$	$X^{(2)}$	$X^{(3)}$	$X^{(4)}$	$X^{(5)}$	$X^{(6)}$	$X^{(7)}$
Réalisations	$X_i^{(1)}$	$X_i^{(2)}$	$X_i^{(3)}$	$X_i^{(4)}$	$X_i^{(5)}$	$X_i^{(6)}$	$X_i^{(7)}$
Valeurs en fonction de i	i	i^2	i^3	i^4	$\sqrt{i}$	$1/i$	$\log i$

Supposons que nous disposons des valeurs prises par 100 réalisations de ces sept variables, c'est-à-dire que l'on considère $i = 1, \ldots, 100$. Ces réalisations de variables qui seront dites potentiellement explicatives seront donc connues. Maintenant, nous allons simuler les réalisations Y_i de la variables à expliquer Y qui ne dépend linéairement que de $X^{(2)}$, $X^{(3)}$ et $X^{(7)}$, c'est-à-dire que l'on simule les réalisations :

$$Y_i = \mu + \beta_2 X_i^{(2)} + \beta_3 X_i^{(3)} + \beta_7 X_i^{(7)} + \varepsilon_i \quad i = 1, \ldots, 100,$$

où $\mu = 5$, $\beta_2 = -0.03$, $\beta_3 = 0.0002$, $\beta_7 = -3$, $\mathrm{Var}(\varepsilon_i) = 20^2$ et ε_i sont i.i.d. de loi gaussienne avec l'espérance nulle.

Voici les commandes en R permettant une telle simulation et traçant également le nuage de points et la vraie fonction pour $i = 1$ à $i = 150$.

```
i = 1:100; X1 = i; X2 = i^2; X3 = i^3; X4 = i^4; X5 = sqrt(i); X6 = 1/i; X7 = log(i);
epsilon = 20*rnorm(100,0); Y = 5-0.03*X2+0.0002*X3-3*X7+epsilon;
j = 1:150; YY = 5-0.03*(j^2)+0.0002*(j^3)-3*log(j);
plot(j, YY,"l", xlim = c(-10, 160), ylim = c(-150, 50)); points(i, Y);
```

On supposera maintenant que sont connues les différentes valeurs Y_i et $X_i^{(1)}, \ldots, X_i^{(7)}$ pour $i = 1, \ldots, 100$. Les ε_i, la variance de ε_i, les coefficients β_j, sont inconnus et surtout on ne sait pas quels sont parmi les coefficients β_j ceux qui sont nuls, c'est-à-dire que l'on ne connaît pas le vrai modèle.

Les commandes suivantes tracent la fonction estimée avec le modèle contenant toutes les variables.

```
y.lm7 = lm(Y~X1+X2+X3+X4+X5+X6+X7);
new = data.frame(X1=j, X2=j^2, X3=j^3, X4=j^4, X5=sqrt(j), X6=1/j, X7=log(j));
y.pred7 = predict(y.lm7, new); points(j, y.pred7, "l", lty="dashed");
```

Pour déterminer le vrai modèle, nous allons appliquer quatre méthodes : régression pas-à-pas descendante, les critères Cp de Mallows, AIC et BIC.

```
# Régression descendante
drop1(lm(Y~X1+X2+X3+X4+X5+X6+X7), test = "F");
...
y.lm.des = lm(Y~?+?+?); y.pred.des = predict(y.lm.des, new);
points(j, y.pred.des, pch = 20);
```

Question 1 : Quelles variables sont retenues après avoir appliqué la régression pas-à-pas descendante ?

```
# Mallows cp
X = matrix(c(X1, X2, X3, X4, X5, X6, X7), ncol = 7);
colnames(X) = c("X1", "X2", "X3", "X4", "X5", "X6", "X7");
library(leaps); r = leaps(X, Y); r$whi; r$Cp;
t = (r$Cp == min(r$Cp)); colnames(X)[r$whi[t]];
y.pred.cp = predict(lm(Y~?+?+?), new); points(j, y.pred.cp, pch = 19, col = "blue");
```

Question 2 : Quelles variables sont retenues après avoir appliqué le critère Cp de Mallows ?

```
# step AIC
step(y.lm7, k = 2); y.pred.aic = predict(lm(Y~?+?+?), new);
points(j, y.pred.aic, pch = 20, col = "red")
# step BIC
step(y.lm7, k = log(100));
```

Question 3 : Quelles variables sont retenues après avoir appliqué les critères AIC et BIC ?

Question 4 : Refaire la simulation est la selection de variables avec les méthodes différentes pour le modèle suivant

$$Y_i = 5 + 0.0002X_i^{(2)} + 10X_i^{(6)} + \varepsilon_i \quad i = 1, \ldots, 100,$$

où $\mathrm{Var}\,(\varepsilon_i) = 0.5^2$ et ε_i sont i.i.d. de loi gaussienne avec l'espérance nulle.

Exercice 2 La chenille processionnaire

Nous allons utiliser un jeu de donnée concernant la chenille processionnaire. On désire connaître l'influence de certaines variables sur la densité de peuplement du parasite. La variable à expliquer est notée Y : c'est le nombre moyen de nids par arbre sur la parcelle considérée de 10 hectares. On dispose ainsi au total des résultats concernant 32 parcelles distinctes. Les différentes variables susceptibles d'avoir une influence sur Y, sont relatives aux différentes caractéristiques de la placette (subdivision de chaque parcelle) et sont :

- l'altitude en mètres : X1 ;
- la pente en degrés : X2 ;
- le nombre de pins dans la placette : X3 ;
- la hauteur (en m) de l'arbre échantillonné au centre de la placette : X4 ;
- le diamètre de cet arbre : X5 ;
- la note de densité de peuplement : X6 ;
- l'orientation de la placette (de 1=sud, à 2=autre) : X7 ;
- la hauteur en mètres des arbres dominants : X8 ;
- le nombre de strates de végétation : X9 ;
- le mélange du peuplement (de 1=mélangé, à 2=non mélangé) : X10.

Il s'agit de données quantitatives même pour X7 ou X10. On va étudier la régression de la variable Y par les variables X1-X10 dans le cadre d'un modèle linéaire. Le but est de choisir parmi les différentes variables explicatives celles qui ont une influence réelle sur la variable Y.

Le fichier de données "proc.txt" est téléchargeable à l'adresse suivante.

http ://samm.univ-paris1.fr/IMG/txt/proc.txt

Enregistrez le dans votre répertoire de travail.