M1 Pro Ingénierie Mathématique — Année 2011-2012
Université d'Angers

TP Régression linéaire

concentration en ozone

Nous allons traiter les données journalières de la concentration en ozone en fonction de la température. Les données se trouvent dans le fichier "ozone.txt". La variable à expliquer est la concentration en ozone, notée "maxO3", et la variable explicative est la température à midi, notée "T12".

1. Commencer par extraire les données grâce à la commande

   ```
   >ozone=read.table("ozone.txt",header=T)
   ```

2. Commencer par représenter les données sur un graphique. Une regression linéaire simple semble-t'elle justifiée graphiquement ?

3. Effectuer la régression linéaire à l'aide de la commande

   ```
   >reg<-lm(maxO3~T12,data=ozone)
   ```

   et consulter les résultats à l'aide de la commande

   ```
   >resume<-summary(reg)
   ```

   Que représente les coefficients de la matrice coefficients ?

4. Tracer l'estimation de la droite de régression, ainsi qu'un intervalle de confiance à 95% de celle-ci grâce aux commandes suivantes :

   ```
   >plot(maxO3~T12,data=ozone)
   >T12=seq(min(ozone[,"T12"]),max(ozone[,"T12"]),length=100)
   >grille<-data.frame(T12)
   >ICdte<-predict(reg,new=grille,interval="confidence",level=0.95)
   >matlines(grille$T12,cbind(ICdte),lty=c(1,2,2),col=1)
   ```

   Ce graphique permet de vérifier visuellement l'ajustement des données au modèle de régression proposé. Que remarquez-vous ?

5. Représentez le vecteur des résidus grâce à la commandes rstudent. Commenter.

6. On s'intéresse à présent à la qualité de prévision du modèle. Pour cela, tracer un intervalle de confiance des prévisions grâce à la fonction predict, en modifiant l'argument interval.

7. On va maintenant calculer les intervalles de confiances des coefficients β_0 et β_1 du modèle de régression. Pour cela, on utilise la fonction coef() qui permet d'extraire les estimateurs de β_0 et β_1 et leurs écarts types empiriques.

   ```
   >seuil<-qt(0.975,df=reg$df.res)
   >beta0min<-coef(resume)[1,1]-seuil*coef(resume)[1,2]
   >beta0max<-coef(resume)[1,1]+seuil*coef(resume)[1,2]
   >beta1min<-coef(resume)[2,1]-seuil*coef(resume)[2,2]
   >beta1max<-coef(resume)[2,1]+seuil*coef(resume)[2,2]
   ```

   Que remarquez-vous sur l'intervalle de confiance de β_0 ? Comment l'expliquez-vous ?

Hauteur des eucalyptus

On veut expliquer la hauteur des eucalyptus en fonction de leur circonférence à partir d'une régression linéaire. On dispose d'observations hauteur-circonférence qui se trouvent dans le fichier "eucalyptus.txt".

1. Extraire et représenter les données dans le plan.

2. Effectuer la régression simple de la hauteur en fonction de la circonférence et commenter les résultats obtenus.

3. Tracer l'estimation de la droite de régression et un intervalle de confiance à 95% de celle-ci. Que déduisez-vous de la qualité de l'estimation ?

4. Calculer les intervalles de confiance des coefficients β_0 et β_1 du modèle de régression et tracer le rectangle de confiance associé.
5. On veut à présent prédire la taille d'une nouvelle série d'eucalyptus de circonférence 50, 100, 200 puis 500. Donner les estimateurs de la taille de chacun d'entre eux et les intervalles de confiances associés. Que se passe-t'il pour les faibles valeurs de circonférences ?
6. Pour améliorer l'estimation, on propose un modèle du type

$$ht = \beta_1 + \beta_2 * circ + \beta_3\sqrt{circ} + \epsilon.$$

   Effectuer la phase d'estimation de cette régression via la formule :

```
>regmult<-lm(ht~circ+I(sqrt(circ)),data=eucalyptus)
```

   L'opérateur I() permet de protéger la racine carrée et sera utilisé à chaque opération sur les variables. Commenter les résultats obtenus.
7. Tracer l'estimation de la droite de régression, ainsi qu'un intervalle de confiance à 95% de celle-ci. Commenter.
8. Tester la significativité du modèle à l'aide du test de Fisher global H_0 : "$\beta = 0 = \beta_1 = \beta_2 = 0$" en utilisant la formule faisant intervenir le R^2. Retrouver le résultat de summary.
9. Tester l'apport de ce modèle de régression multiple par rapport au modèle de régression simple à l'aide d'un test emboîté H_0 : "$ht = \beta_0 + \beta_1 * circ$" contre H_1 : "$ht = \beta_0 + \beta_1 * circ + \beta_2\sqrt{circ}$" grâce à la fonction anova.
10. Retrouver le résultat dans la matrice coefficients.

Consommation de glace

On étudie la consommation de glace aux Etats-Unis sur une période de 30 semaines du 18 Mars 1950 to 11 Juillet 1953. Les variables sont la période (de la semaine 1 à la semaine 30), la consommation (Consumption en pintes par habitant), le prix des glaces (Price en dollars), le salaire hebdomadaire (Income en dollars), et la température (Temp en degré fahrenheit). Les données sont disponibles dans le fichier "icecream-R.dat".

1. Extraire les données et représenter la consommation en fonction des différentes variables. Représenter l'évolution du salaire (Income) en fonction de la période. Interpréter.
2. On propose de régresser la consommation sur les trois variables Price, Income et Temp. Réaliser la phase d'estimation de cette régression et commenter les résultats obtenus.
3. Déterminer les intervalles de confiance simultanés au niveau au moins 95% pour les β_j, $j = 0, \dots 3$ par la méthode de Bonferroni.
4. Construire les régions de confiance des couples (β_i, β_j) de paramètres et les comparer graphiquement aux intervalles de confiance grâce aux commandes suivantes :

```
>library(ellipse)
>plot(ellipse(regmult,c(i+1,j+1),level=0.95,type="l",xlab=paste("beta",i,sep=""),
ylab=paste("beta",j,sep="")))
>points(coef(resume)[i],coef(resume)[j],pch=3)
>IC<-rbind(coef(resume)[,1]-coef(resume)[,2]*qt(0.975,regmult$df.res),coef(resume)
[,1]+qt(0.975,regmult$df.res))
>lines(c(IC[1,i],IC[1,i],IC[2,i],IC[2,i],IC[1,i]),c(IC[1,j],IC[2,j],IC[2,j],
IC[1,j],IC[1,j]),lty=2)
```

   Qu'apporte comme information supplémentaire ces ellipses de confiance ?
5. Tester la significativité du modèle proposé à l'aide du test de Fisher global : H_0 : "$\beta_2 = \beta_3 = \beta_4 = 0$".
6. Tester H_0 : "$Intercept = 0$" puis H_0 : "$Price = 0$". Tester à l'aide de la fonction anova le modèle (sans constante) réduit aux variables "Income" et "Temp". Commenter.
7. Retrouver ces résultats à l'aide de la fonction linear.hypothesis, qui permet de faire des tests d'hypothèses linéaires (documentation disponible à l'adresse suivante : www.math.univ-angers.fr/ loustau).

8. Tester à l'aide de la fonction linear.hypothesis H_0 : "$Income = Temp$".
9. On s'intéresse à la prédiction de consommation de nouvelles données. Déterminer l'estimation ponctuelle $\hat{y}$ et l'intervalle de confiance associé à chacune des données suivantes :
   - x_1=(Price=0.3, Income=85, Temp=65) ;
   - x_2=(Price=0.26, Income=76, Temp=71) ;
   - x_3=(Price=0.26, Income=85, Temp=90).
10. Déterminer par la méthode de Scheffé les intervalles de confiances simultanés de $y(x_i)$, $i = 1, 2, 3$.
11. Régresser la consommation sur le salaire et la température dans un modèle sans constante. Estimer les paramètres de la régression et répéter les questions 9. et 10. Commenter.

Effet du vent sur la concentration en ozone

On veut étudier l'effet de la direction du vent sur les pics d'ozone. Pour cela, on va considérer le fichier "ozone.txt" et expliquer la variable "maxO3" par la variable qualitative "vent".

1. Importer les données et résumer les variables d'intérêts, ici "maxO3" et "vent" :

   ```
   >ozone<-read.table("ozone.txt",header=T)
   >summary(ozone[,c("maxO3","vent)])
   ```

2. Représenter les données à l'aide des boîtes à moustaches pour illustrer l'effet du vent sur les pics d'ozone :

   ```
   >plot(maxO3~vent,data=ozone,pch=15,cex=.5)
   >summary(ozone[,c("maxO3","vent)])
   ```

   Commenter.

3. Réaliser l'analyse de variance pour estimer les paramètres du modèle :

   ```
   >regaov<-lm(maxO3~vent,data=ozone)
   >summary(regaov)
   ```

   Que représente la ligne Intercept et la colonne Estimate ? Quelle contrainte a-t'on implicitement imposer sur les paramètres ?

4. Retrouver ces résultats à la main en calculant dans ce cas $(X^{*\prime}X^*)^{-1}$.
5. Tester à présent la significativité du modèle, à l'aide du tableau d'analyse de variance :

   ```
   >anova(regaov)
   ```

   Conclure.

6. On veut à présent imposer la contrainte $\mu = 0$ d'effet moyen nul. Pour cela, il suffit de spécifier au logiciel un modèle sans constante :

   ```
   >regaov2<-lm(maxO3~-1+vent,data=ozone)
   >summary(regaov2)
   ```

   Que représente dans ce cas la collone Estimate de la matrice Coefficient. Retrouver les valeurs des estimateurs "à la main".

7. On veut tester l'influence du vent sur le pic d'ozone grâce à un tableau d'analyse de variance. On propose d'utiliser la commande suivante

   ```
   >anova(regaov2)
   ```

   Ce tableau d'analyse de variance est faux. Quand la constante ne fait pas partie du modèle, tester $H_0 : \alpha_1 = \ldots = \alpha_I = 0$ n'a pas de sens pour illustrer l'effet du facteur.

8. Pour des raisons particulières, on peut choisir une cellule témoin spécifique (R choisit par défaut la première par ordre alphabétique, ici Est) avec la commande suivante :

   ```
   >regaov3<-lm(maxO3~C(vent,base=2),data=ozone
   ```

   Retrouver les résultats de la première analyse de variance, seul l'ordre des coefficients étant modifié.

9. On peut aussi choisir la contrainte $\sum_{i=1}^{p} \alpha_i = 0$ grâce à la commande :

   ```
   >regaov3<-lm(maxO3~C(vent,sum),data=ozone)
   ```

   Interpréter les entrées de la matrice Coefficients dans ce cas. Comment estimer l'effet du vent du Sud ?