ENSC 2A - "Modélisation statistique"

TP 4 : Analyse de la covariance (ANCOVA) avec le logiciel R

1 Un premier exemple de données réelles

On observe, pour deux niveaux de pression (facteur `pression`), les valeurs `X` et `Y` qui sont respectivement le pourcentage de molibdène et le niveau de densité de l'acier obtenu. Les ingénieurs travaillant sur la fabrication de l'acier s'interrogent sur l'influence ou pas du facteur `pression` et de la variable quantitative `X` sur la densité moyenne de l'acier. Pour cela, il décide de mettre en œuvre une analyse de covariance sur les données d'exprimentation qui sont disponibles dans le "data frame" `acier2.Rda`.
Le traitement statistique a été fait avec le logiciel R (voir les lignes de codes dans la section suivante). Au vu des résultats obtenus, quelle conclusion donneriez-vous à cette étude ?

1.1 Mise en œuvre avec R

```
# Importation des donnees
#--------------------
load("acier.Rda")
head(acier)

class(acier$densite.Y)
class(acier$molibdene.X)
class(acier$pression)
table(acier$pression)

# Quelques petits graphiques utiles
#-------------------------------

par(mfrow=c(1,2))
plot(densite.Y~ pression,data=acier)
plot(acier$molibdene.X,acier$densite.Y)
par(mfrow=c(1,1))
couleur<-c(rep(1,25),rep(2,25))
plot(acier$molibdene.X,acier$densite.Y,col=couleur)

# Quelques statistiques descriptives
#-----------------------------------------------
summary(acier)
tapply(acier$densite.Y,acier$pression,summary) # par groupe de pression

# Recherche du "meilleur" modele
#--------------------------------
res1 <- lm(densite.Y~molibdene.X*pression,data=acier)   # modele complet (avec interaction)
anova(res1)
summary(res1)

res2 <- lm(densite.Y~molibdene.X+pression,data=acier)    # modele sans interaction
anova(res2)
summary(res2)

# Etude des residus et de l'homoscedasticite (modele 2)
#-------------------------------------------------------

shapiro.test(res2$residuals)  # test de normalite des residus
bartlett.test(acier$densite.Y,acier$pression)     # test d'homoscedasticite
plot(res2$fitted,res2$residuals)     # graphique des valeurs predites versus les residus
abline(h=0,col=2)
```

2 Autres jeux de données

2.1 Données sur le traitement de la lèpre

Les données concernent sur l'usage de médicaments dans le traitement de la lèpre. Les variables de l'étude sont :

- `drug` = variable qualitative à 3 modalités : deux antibiotiques (A et D) et un non-traitement (F pour avoir un groupe de contrôle),
- `X` = score de pré-traitement du bacille de la lèpre,
- `Y` = score de post-traitment du bacille de la lèpre.

Dix patients ont été sélectionnés pour chaque niveau du facteur `drug`, et six sites sur chaque patient ont été mesurés pour le bacille de la lèpre. La covariable `X` (score de pré-traitement) a été incluse dans le modèle afin d'augmenter la précision pour déterminer l'effet du facteur DRUG sur le score de post-traitement de ce bacille.

Les données disponibles sont contenues dans le "data frame" `lepre.Rda`.
Faire l'analyse de la covariance correspondant à la problématique posée.

2.2 Retour sur les données "ozone" du TP 2

Dans le cadre du TP 2 sur la régression linéaire multiple, on a cherché à expliquer la variable `max03` (maximum journalier de la concentration en ozone (en $\mu g/m^3$)) en fonction des autres variables quantitatives disponibles : des variables de température `T9, T12, T15`, des variables de nébulosité `Ne9, Ne12, Ne15`, des variables de vent `Vx9, Vx12, Vx15`, et aussi de la mesure du maximum de la concentration en ozone de la veille `max03v`.

Le jeu de données "ozone" (fichier texte `ozone.txt`) contient aussi deux variables qualitatives : le facteur `vent` indiquant la direction du vent et le facteur `pluie` indiquant si le temps est pluvieux ou sec.

Travail à réaliser.

- Intégrer dans la modélisation ces deux facteurs en plus des variables quantitatives disponibles et proposer le "meilleur modèle". Commenter.
- Faire une classification de variables et regarder à quelles classes appartiennent les facteurs `vent` et `pluie`. Commenter.

  NB : pour cela, utiliser le package R `ClustOfVar`.