TD 1 : Régression Linéaire avec R

1 Calcul des coefficients de la droite de régression

On suppose qu'on dispose de n données (x_i, y_i) contenues dans un fichier .csv. On commence par lire ce fichier en affectant les valeurs dans une data frame nommées ici "donnees" puis on trace le nuage de points pour vérifier visuellement la potentielle linéarité.

```
> # Lecture des données dans un fichier csv (séparateur point virgule)
> donnees <- read.csv2("Droite.csv")
> print(t(donnees))# x=variable explicative,y=variable expliquée
> # t pour transposer le tableau
> # pour accèder aux variables de donnees par leur nom (seul) (simplifier leur nom)
> attach(donnees)# par ex donnees$x devient x
> # tracé: nuage de points
> eq = paste0("Nuage de points: observations")
> plot(x, y,type="p",pch=4,main=eq,col="blue")
```

Les points semblent alignés, on espère donc que le modèle suivi par les données est une droite d'équation

$$y = a_1 + a_2 x$$

dont il faut déterminer les coefficients.

1.1 Calcul matriciel

La méthode des moindres carrés consiste à déterminer les valeurs a_1 et a_2 qui minimisent les écarts au carré entre les valeurs expérimentales et les valeurs prédites par le modèle (écarts aussi appelés résidus). C'est un problème de recherche du minimum d'une fonction de deux variables a_1 et a_2 (cf cours):

$$S(a_1, a_2) = \sum_{i=1}^{n} (a_1 + a_2 x_i - y_i)^2$$

Cela conduit à un système linéaire qu'il suffit d'inverser

```
> # y=a1+a2*x détermination de a1 et a2 par la méthode des moindres carrés
> # Construction de X et Y (cf cours)
> c1=rep(1,10)
> X<-matrix(c(c1,donnees$x),ncol = 2)
> Y<-matrix(donnees$y,ncol=1)
> M=t(X)%*%X # transposée(X) *X
> b<-t(X)%*%Y # transposée(X) *Y
> a=solve(M,b) # inversion du système donne le vecteur a=les coefficients a1 et a2
> plot(x, y,type="p",pch=4,main=eq,col="blue")# tracé du nuage de points (x,y)
> abline(a, col="red",main=eq)# tracé droite de regression (courbe de tendance linéaire)
> # equation de la droite sous forme de chaine de caractères
> eq = paste0("Droite regression: y = ", round(a[2],3), "*x +",round(a[1],3))
> title(sub=eq)# ajout d'un sous titre contenant l'équation de la droite
```

Nous avons ainsi déterminé la droite de régression par la méthode des moindres carrés c'est-à-dire en minimisant les écarts au carré entre les points expérimentaux et la droite.

1.2 Coefficients de la droite de régression avec lm

Il existe sous R une fonction qui permet le calcul des coefficients de la droite de régression sans avoir besoin de programmer les matrices X et Y (la fonction le fait elle même). C'est la fonction lm (pour Linear Models).

```
> droite <- lm(y~x, data=donnees)
> # y~x signifie chercher une relation entre y et x du type y=a1 + a2*x sur les données de la data frame donn
> print(droite) # affichage du résultat nommé ici droite qui est une liste
> a1 <- coef(droite)[1] # ordonnée à l'origine appelé intercept
> a2 <- coef(droite)[2] # la pente
> # tracé: nuage de points avec un titre principal et en sous titre l'équation
> eq = paste0("y = ", round(a2,3), "*x +",round(a1,3))
> plot(x, y,main="Droite de régression", sub=eq)
> # tracé: droite de regression (courbe de tendance linéaire)
> abline(droite, col="red")
```

On retrouve bien les mêmes coefficients qu'avec le calcul matriciel (évidemment, c'est le même calcul !). Mais on obtient en plus d'autres résultats.

2 Approche probabiliste: qualité de la prédiction

Si on veut estimer la qualité de la prédiction, on peut avoir une approche probabiliste: elle consiste à faire l'hypothèse que dans les observations notées (x_i, y_i), les x_i sont des valeurs exactes et certaines d'une variable non aléatoire x et les y_i sont des réalisations d'une variable aléatoire notée Y.
On fait la supposition que cette V.A. vérifie

$$Y = \alpha_1 + \alpha_2 x + \epsilon$$

où l'erreur ϵ est une variable aléatoire vérifiant $\epsilon \sim \mathcal{N}(0, \sigma^2)$. On peut montrer que les paramètres, a_1 et a_2, déterminés par les moindres carrés, sont alors les meilleurs estimateurs sans biais de α_1 et α_2 [1] de la méthode de maximum de vraisemblance. On peut utiliser cette régression pour répondre à certaines questions statistiques (signification, comparaison, prédiction, ...) que nous allons détailler. Mais on doit vérifier que $\epsilon \sim \mathcal{N}(0, \sigma)$ en étudiant la répartition des réalisations de la variable aléatoire, les résidus r_i (où $r_i = a_1 + a_2 x_i - y_i$).

2.1 Qualité de la prédiction: R^2, résidus, écart-type, tests...

2.1.1 Analyse de la variance

Lors de l'appel de lm, l'analyse de la variance a été faite

```
>   anova(droite)# table d'analyse de la variance
> # Response: y
> #           Df   Sum Sq  Mean Sq       F value   Pr(>F)
> # x         Df1  SSR     MS1=SSR/Df1   MS1/MS2    pvalue prob(F(alpha,Df1,Df2)>F)
> # Residuals Df2   SSE    MS2=SSE/Df2
> # H0 "pente nulle" est rejetée si pvalue<risque_alpha
>
> # SSR=sum square regression (variance expliquée par le modèle)
> # SSE= sum square error (variance non expliquée par le modèle)
> # SST=SSE+SSR (variance totale)
```

Dans la table, on obtient en colonnes

- les dégrés de liberté "Df" (pour les lois des tests sur les paramètres)
- les sommes des carrés (Sum Sq): Sum Square Regression: variance expliquée par le modèle et Sum Square Erreur: variance résiduelle non expliquée par le modèle.
- les moyennes de sommes des carrés (Mean Sq): SSR/Df1 et SSE/Df2 (estimation de la variance σ^2 de l'erreur ϵ)
- la statistique (F value) (quotient de deux valeurs précédentes-suit une distribution de Fisher $\mathcal{F}_{1,n-2}$).
- la p-value (Pr(>F)) associée au test d'hypothèse H_0 : "la pente a_2 est nulle" contre H_1 : "la pente n'est pas nulle". On rejette H_0 au risque α, si la p-value est inférieure à α. Dans notre cas, par exemple, on rejette l'hypothèse "la pente est nulle".

[1] Plus loin, on notera α le risque (à ne pas confondre avec α_1 et α_2)

2.1.2 Commande Summary

On peut obtenir d'autres informations statistiques, en tapant la commande summary().

```
> summary(droite)# résumé des calculs effectués par l'instruction lm
> # Coefficients:
> # Estimation a1, sa1=ecart-type de a1, t=a1/sa1, probabilité critique PCa1
> # H0: la droite passe par l'origine"vs H1 la droite ne passe pas pas l'origine
> # H0 est rejetée si t<student(n-2,1-risque/2) ie si la pvalue PCa1<risque.
> # Estimation a2, sa2=ecart-type de a, t=a2/sa2, probabilité critique PCa2
> # H0: la pente de la droite est nulle" rejetée si t<student(n-2,1-risque/2)
> # H0: la pente de la droite est nulle" rejetée si la pvalue PCa2<risque.
> # Residual standard error: SQRT(MS2)
> # doit être faible pour que le modèle soit considéré comme bon (=prédictif)
> # Multiple R-squared:  SSR/SST,
> # Plus cette valeure sera proche de 1 meilleur sera l'ajustement.
> # Adjusted R-squared:  ajustement du R^2 au nombre p de variables explicatives
> #F-statistic:  MS1/MS2 on Df1 and Df2,  p-value
```

Dans la 1ère table "Residuals", on trouve les minimum et maximum ainsi que les quartiles des résidus. On peut aussi afficher les valeurs des résidus avec la commande

```
> resid(droite)# les résidus
```

Dans la seconde table "Coefficients"

- la colonne "estimate" contient les estimations de l'ordonnée à l'origine, a_1 appelé (intercept) puis de la pente, a_2,
- leurs écarts-types (Std. Error) respectifs,
- pour chaque paramètre, la statistique observée (t value) ainsi que la p-value (Pr(>|t|)) associée au test d'hypothèse H_0 : "le paramètre est nul" contre H_1 : "le paramètre n'est pas nul". On rejette H_0 au risque α, si la p-value est inférieure à α. Dans notre cas, par exemple, on rejette les hypothèses "la pente est nulle" et "la droite passe par l'origine".

Ensuite, on trouve

- "Residual standard error" l'estimation s_r de l'écart-type σ de la VA ϵ où $s_r = \sqrt{\frac{SSE}{Df2}}$.
- "Multiple R-squared" le coefficient de détermination $R^2 = \frac{SSR}{SST}$ qui exprime le rapport entre la variance expliquée par le modèle et la variance totale. Ce coefficient varie entre 0 et 1. S'il est égal à 1, cela signifie que le modèle explique parfaitement les données.
- "Adjusted R-squared" le coefficient de détermination ajusté (prend en compte le nombre de variables)
- la statistique F déjà définie dans l'anova.

2.1.3 Analyse des résidus

Une étape fondamentale dans la démarche de la régression est l'étude des résidus. En effet, les résultats statistiques donnés précédemment sont basés sur l'hypothèse que les erreurs (les résidus) sont indépendantes, normalement distribuées, de moyenne nulle et de variance constante. La fonction lm fournit quatre graphiques qui permettent d'en juger.

```
> layout(matrix(1:4,2,2))# fenetre graphique coupée en 4
> plot(droite) # 4 graphiques
```

Le premier graphe "Residual vs fitted" donne les résidus en fonction des valeurs prédites. Les points doivent être répartis aléatoirement autour de l'axe horizontal $y = 0$ et ne pas montrer de tendance. Le graphe "Scale-Location" donne la racine des résidus standardisés en fonction des valeurs prédites et ne doit pas non plus montrer de tendance. Le graphe "Normal Q-Q" permet de vérifier la normalité des résidus en comparant les quantiles de la population avec ceux de la loi normale. Le dernier graphique "Residuals vs Leverage" met en valeur l'importance de chaque point dans la regression. On se questionnera particulièrement sur les points ayant une distance de Cook supérieure à 1 (rend la donnée suspecte=>point aberrant ?).
On peut aussi afficher les résidus studentisés qui en pratique devront être compris entre les bornes -2 et 2.

```
> # résidus studentisés
> res <- rstudent( droite)
> # tracé des résidus studentisés (ylim échelle pour y)
> plot(res,ylab="Résidus studentisés",xlab="",main="Résidus de student",ylim=c(-2.5,2.5))
> # tracés des droites y=0 (trait plein bleu), y=+/- 2 (trait espacé rouge)
> # h= contient équation y=(-2,0,2) et lty le type de lignes
> abline(h=c(-2,0,2),lty=c(2,1,2),col=c("red","blue","red"))
```

En conclusion, les points suspects sont les points dont le résidu studentisé est supérieur à 2 en valeur absolue et/ou la distance de Cook est supérieure à 1. Dans ce dernier cas, le point contribue très/trop fortement à la détermination des coefficients du modèle comparativement aux autres. Il n'y a pas de méthode universelle pour traiter ce type de points.

2.1.4 Les prédictions: intervalle de confiance et de prédiction

On peut obtenir les valeurs prédites par la commande fitted

```
> fitted(droite)# les valeurs prédites
```

On peut obtenir de nouvelles prédictions avec intervalle de confiance.

```
> newx=seq(1.5,10,1);
> pc=predict(droite,data.frame(x= newx), level = 0.95, interval = "confidence")
> print(pc)
```

Attention, un intervalle de confiance n'est pas un intervalle dans lequel la valeur a une probabilité de 95 % de se trouver. L'IC a 95 % de chance de contenir la vraie valeur si on répète les estimations un grand nombre de fois. Autrement dit, un intervalle de confiance à 95 % donnera un encadrement correct 95 fois sur 100.

On peut aussi obtenir un intervalle de prédiction

```
> newx=seq(1.5,10,1);
> pp=predict(droite,data.frame(x= newx), level = 0.95, interval = "prediction")
> print(pp)
```

Un intervalle de prédiction est un intervalle qui est susceptible de contenir une observation individuelle future. L'intervalle de prédiction est toujours plus large que l'intervalle de confiance.

On peut afficher les prédictions, les bandes de confiance et de prédiction

```
> plot( pp[,1] ~ newx, type='p',pch=3,ylab="prédictions",xlab="x" )
> points( pc[,2] ~ newx, type='l', col="green" )
> points( pc[,3] ~ newx, type='l', col="green" )
> points( pp[,2] ~ newx, type='l', col="red" )
> points( pp[,3] ~ newx, type='l', col="red" )
> title(main="Bandes de confiance et de prédiction")
> legend("topleft", c("Bande de confiance", "Bande de prédiction"),lwd=1, lty=1, col=c("green", "red") )
```

3 Généralisation

On peut faire de la régression linéaire autre que le cas de la droite avec R. La syntaxe est lm(y∼formule)
où y est le nom de la variable à prédire et formule est l'équation du modèle. Pour accèder aux différents calculs statistiques, il faut définir une liste contenant le résultat res=lm(y∼formule)

Formule	Modèle	Commentaires
y~x	$y(x) = a_1 + a_2 x$	droite
y~-1+ x	$y(x) = a_2 x$	droite passant par l'origine
y~x+I(x^2)	$y(x) = a_1 + a_2 x + a_3 x^2$	polynôme d'ordre deux. La fonction I() permet de programmer des fonctions mathématiques
y~poly(x,deg)	$y(x) = \sum_{i=0}^{deg} P_i(x)$	polynômes (orthogonaux) de Legendre d'ordre deg.
y~x+z	$y(x) = a_1 + a_2 x + a_3 z$	termes du 1er ordre en x et z sans terme en xz
y~x:z	$y(x) = a_1 + a_2 xz$	1er ordre seulement le terme en xz
y~x*z	$y(x) = a_1 + a_2 x + a_3 z + a_4 xz$	1er ordre complet
y~(x+z+t)^2	$y(x) = a_1 + a_2 x + a_3 z + a_4 t + a_5 xz + a_6 xt + a_7 zt$	équivalent à x*z*t-x:z:t

4 Exercice

Exercice 4.1 *On dispose des données expérimentales dans le fichier poly.csv (à télécharger sour Moodle).*

La commande **nom=file.choose()** *permet de choisir un fichier dans le répertoire. En utilisant cette commande, sélectionner le fichier poly.csv*

(b) *Lire les données contenues dans nom (poly.csv) en les affectant à une data frame nommée obs et on fera un attach(obs) pour simplifier le nom des variables.*

(c) Afficher le nuage de points ($conc_i$,DO_i).

(d) Détermier les paramètres (α_1, α_2) du modèle

$$DO = \alpha_1 + \alpha_2 conc + e$$

(e) Tracer sur un même graphe le nuage de points et la courbe théorique obtenue en rouge. Ajouter un titre. D'autres commandes parfois utiles

(f) Etudier les résidus (residus en fonction des valeurs prédites, résidus standardisés, Distance de Cook, résidus studentisés, ...) et commenter.

(g) Détermier le paramètre α du modèle

$$DO = \alpha_2 conc + e$$

On cherche maintenant à approcher le nuage de points selon le critère des moindres carrés par un polynôme de degré deux $\alpha_1 + \alpha_2 conc + \alpha_3 conc^2$.

(d) Détermier les paramètres $(\alpha_1, \alpha_2, \alpha_3)$ en utilisant la fonction lm de Rstudio.

(b) Tracer sur un même graphe le nuage de points et la courbe théorique obtenue (cf help: curve) en rouge. Ajouter un titre.

(c) Etudier les résidus comme précédemment.

2. *Quel est le meilleur modèle ? On peut regarder les paramètres et leurs écart-type, les résidus, le R^2_{adj}, utiliser le critère AIC (plus le critère est faible, meilleur est le fit).*