TDM : Régression

Concernant les TDM ou TP qui ont lieu en B141, vous utiliserez l'image **Deb-NF-MA431-2021-img**. Vous commencez par ouvrir Rstudio, créez un nouveau projet dans lequel **vous déposerez les données**. Créez ensuite un fichier RMarkdown en précisant une sortie en pdf ou en html. Vous enregistrerez vos commandes dans ce R Markdown en respectant la syntaxe ainsi que vos commentaires entre les différentes commandes. Pour compiler el fichier R Markdown, vous devez utilisez le bouton knit (tricoter).
Rstudio a l'avantage d'être composé de quatre consoles, une, en haut à gauche, pour les R scripts ou R MArkdown, une pour l'environnement et l'historique (en haut à droite, avec les différentes variables, données, etc.), une pour les résultats de vos commandes validées (en bas à gauche) et une pour l'aide, les packages et les graphiques notamment (en bas à droite).

Dans la suite de l'énoncé, toutes les commandes R seront indiquées en *italique*.
Vous devez répondre en commentaire soit en dehors de la zone de commandes R, soit sur la ligne de la commande en utilisant le #.
Vous pouvez vous inspirer de la syntaxe proposée lorsque vous ouvrez un fichier RMarkdown. Les titres sont précédés de ## et les zones de commandes R sont entre "'{r} et "'.

1 Régression linéaire

Exercice 1

Récupérer sur Chamilo le fichier TDM_Ex_Reglin.csv et le sauvegarder dans le dossier qui contient votre projet R.

A. **Etude des données**

(a) Importez ensuite les données dans R à l'aide de la fonction *read.csv2* et stockez les dans une variables données.
Quelles sont les données manipulées et leur nature ? Combien de variables observe-t-on ? sur combien d'individus ? (vous pouvez utiliser les fonctions *dim* et *str*)

(b) Une variable est inutile. Laquelle ? Supprimez-la à l'aide de la commande *donnees← données[,-numéro de la colonne]*.

(c) Représentez graphiquement les données simultanément avec *pairs(donnees)*. Pensez-vous qu'il y ait des régressions linéaires possibles ? Quelle régression privilégierez-vous ?

B. **Régression linéaire**

(a) Affecter les variables de données à y, x_1 et x_2. Effectuez la regression linéaire de y en fonction de x1. Affichez les coefficients et les résidus puis interprétez ces résultats.

(b) Afficher un résumé des informations de la régression avec la commande *summary*. Interprétez précisément les résultats.

(c) Représentez graphiquement le nuage de points et la droite de régression sur le même graphique. Impression ?

(d) Rappelez les **hypothèses à vérifier sur les résidus**.

(e) Pour vérifier une de ces hypothèses, calculez la moyenne des résidus. Conclusion ?

(f) Pour une autre hypothèse, on va tracer un diagramme Quantiles-Quantiles appelé Droite de Henry. *qqnorm(reg1$residuals,xlab="Quantiles théoriques",ylab="Quantiles observés")* . Pour observer la normalité de la distribution des résidus, on compare la distribution des résidus à celle d'une loi normale. Si les points sont approximativement alignés, la distribution est compatible avec une loi normale. Ce test est une vérification empirique, non une preuve statistique. Quelle conclusion peut-on en déduire avec les résidus obtenus ?

C. **Autres régressions linéaires**
Il reste au moins deux autres régressions à déterminer. Effectuer la regression de y en fonction de x_2 puis de y en fonction de x_1 et x_2. Déterminer **le meilleur modèle parmi les trois**. Justifier.

2 Régression logistique

Exercice 2

Créer un nouveau fichier RMarkdown et le sauvegarder au nom de TDMReglog. Récupérer sur Chamilo le fichier TDM_Ex_Reglog.csv et le sauvegarder dans le dossier qui contient votre projet R.
On dispose d'un échantillon de mails sur lesquels on a observé des variables puis que l'on a identifié en SPAMS (1) et NONSPAMS (0). L'objectif est d'obtenir un moyen de classer des mails en SPAMS à l'aide d'une regression logistique.

1. Importez les données et stockez-les dans une variable SPAMS. Que contient la base de données SPAMS ? Combien d'individus et de variables ?
2. Effectuez un résumé des données ainsi que des représentations graphiques croisées puis interprétez.
3. Affecter les variables de la base de données à des variables que nous noterons b pour la variable à expliquer et vi pour la ième variable explicative. **Effectuez une regression logistique** de b en fonction de toutes les vi. Affichez un résumé des coefficients puis interprétez.
4. Effectuez un test sur la différence entre la déviance nulle et la déviance résiduelle (la différence de ces deux déviances suit une loi du χ^2 avec pour degrés de liberté la différence des deux degrés de libertés). Conclusion du test ?
5. On observe sur un message $x1 = 8, x2 = 10$ et $x3 = 45$. D'après le modèle, ce message est-il un SPAM ? Vous écrirez le modèle puis utiliserez la fonction suivante
   predict.glm(rlog123,newdata=data.frame(v1=8,v2=10,v3=45),type="response") afin de répondre à la question. La fonction predict.glm renvoie directement la probabilité d'obtenir 1 avec les nouvelles données.
6. Construitre la matrice de confusion puis interprétez. Vous pourrez calculer différents taux.
7. **Construction de la courbe ROC :**
   Y=b s=0
   plot(0 :1,0 :1,xlab="Taux de Faux Positifs", ylab="Taux de Vrais Positifs",cex=.5)
   for(s in seq(0,1,by=.01)){
   *Ps=(score>s)*1*
   FP=sum((Ps==1)(Y==0))/sum(Y==0)*
   VP=sum((Ps==1)(Y==1))/sum(Y==1)*
   points(FP,VP,cex=.5,col="red")
   } abline(a=0,b=1,col="green")
   Interprétez la courbe ROC obtenue.
8. Pour tester les modèles logistiques emboités, nous allons utiliser une fonction de la librairie MASS. Il faut donc commencer par l'activer : *library(MASS).*
   La fonction stepAIC permet de tester, sur le critère de l'AIC, tous les modèles emboités en partant du premier proposé jusqu'au dernier et en choisissant la direction (backward ou forward). La fonction s'arrête au meilleur modèle emboité.
   pasapas← stepAIC(rlog123,scope=list(lower="b∼ v1",upper="b∼ v1+v2+v3"),direction="backward")
   Interprétez les résultats.
9. Recommencer mais en partant de "b∼ v3". Conclusion ?