Sujet de Travail Dirigé Master 2 SISE

2017/2018

TD – SCILAB Niveau 2

I. Régression linéaire simple :

On a un échantillon de 10 individus d'âges différents, on a recueilli pour chacun d'eux leur concentration sanguine en cholestérol (g/L) :

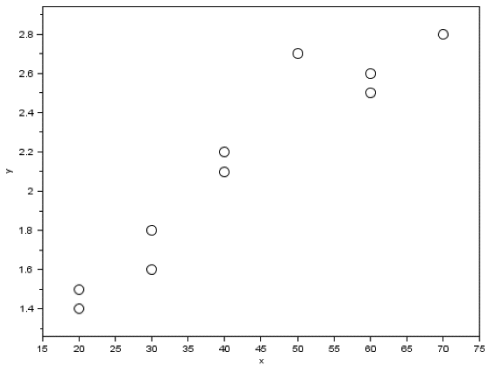
Age (xi)	30	60	40	20	50	30	40	20	70	60
g/L (yi)	1.6	2.5	2.2	1.4	2.7	1.8	2.1	1.5	2.8	2.6

⇨ On cherche à expliquer la concentration sanguine en cholestérol à partir de l'âge des individus.

Aide SCILAB : https://help.scilab.org/docs/5.5.2/fr_FR/index.html

http://enseignement.math.univ-angers.fr/documents/divers/documentation_informatique/mementoScilab.pdf

1) Saisir ces données dans Scilab sous la forme de deux vecteurs.
2) Pour chacun de ces vecteurs, stocker dans une variable le minimum, le maximum, la moyenne, la variance et l'écart type. Calculer la covariance de x et y.
3) Représenter graphiquement les individus à l'aide d'un nuage de points (plot2D) et déterminer si les variables sont liées entre elles.



4) Calculer les indicateurs suivants :

- Somme des carrés totale (SCT)
- Somme des carrés résiduelle (SCR)
- Somme des carrés expliquée (SCE)
- La valeur de r^2

5) La corrélation est-elle significative à 5% ?
6) Tester la significativité globale de ce modèle à 95%. Ces résultats sont-ils en accords avec les hypothèses réalisées à partir du nuage de points ?

Source des données : http://www.agrocampus-ouest.fr/math/livreR/ozone.txt

(voir table : http://maths.cnam.fr/IMG/pdf/table_Fisher-Snedecor.pdf)

7) Pour une personne de 45 ans, quelle serait la prédiction du modèle ?
8) Représenter la droite de régression dans le graphique.

I. Régression linéaire multiple :

Un relevé de la concentration de certains composants chimiques dans l'air, comme le dioxyde d'azote (NO2) et l'ozone (O3), a été effectué à Lyon en mars et avril 2014. Elles enregistrent également les conditions météorologiques comme la température, la couverture nuageuse, le vent, etc.

La variables endogène (Y) :

NO2 : maximum journalier de la concentration en dioxyde d'azote (en _g/m3) ;

Les variables exogènes :(X1,X2….. Xn) :

O3 : maximum journalier de la concentration en ozone (en _g/m3) ;
Tmin : température minimale journalière (en_C) ;
Tmax : température maximale journalière (en_C) ;
Tmoy : température moyenne journalière (en_C) ;
VentMax : vitesse du vent (en Km/h) ;

1. Chargez le fichier « base_ozone.txt» [cf. **csvRead()** & pXX ; le séparateur de colonnes est le caractère tabulation « \t », décimal en « . » et les données sont importés en texte] la concentration du NO2 dans la variable y, les variables explicatives vont de x1,…, x5. Utiliser la fonction **strtod** pour convertir les données en numérique et **ones** pour crée la matrice X.
   Quelques indications :
   http://www.geoazur.fr/PERSO/hassani/RiadHassani/Enseignement_Teaching_files/memento-scilab.pdf
   https://help.scilab.org/docs/5.4.1/fr_FR/csvRead.html

2. Réalisez une régression linéaire multiple expliquant la variable NO2 à partir de toutes les autres (https://www.math.univ-toulouse.fr/~besse/Wikistat/pdf/st-l-inf-intRegmult.pdf p2
3. Calculez l'estimation des paramètres βj en utilisant la fonction **inv().** Calculez ensuite les valeurs ajustées dans y^.
4. Récupérez les résidus de la régression. Calculez sa moyenne. Que constatez-vous ?

5. Calculez la somme des carrés des résidus SCR (https://www.math.univ-toulouse.fr/~besse/Wikistat/pdf/st-l-inf-intRegmult.pdf p2 3.3). pour calculer e^2 il sera nécessaire de transposer l'une des matrices avec '. Calculez ensuite la somme totale des carrés SCT et la somme des carrés expliqué SCE. Calculez les valeurs de R^2.

(https://eric.univ-lyon2.fr/~ricco/cours/cours/Regression_Lineaire_Multiple.pdf p22 et p24)7

6. La régression est-elle globalement significative ? (cf lien ci-dessus p22) .

Source des données : http://www.agrocampus-ouest.fr/math/livreR/ozone.txt

7. Donnez une prédiction de la concentration maximum du NO2 pour une journée avec une concentration max en O3 de 37g/m3, une température min de 3C, une température max de 9c, une température moyenne de 5C et une vitesse de vent de 17km/h .

II. Analyse discriminante prédictive :

Modèle prédictif de faible poids de naissance d'un bébé.

But: cette étude a pour objectif de mettre sur pied une modélisation de prédiction du poids de naissance à travers les caractéristiques maternelles susceptibles d'influencer ce dernier.

Variable à prédire :

- Faible poids à la naissance 1 = BWT <= 2500g, FAIBLE 0 = BWT> 2500g

Les variable prédictives :

- Âge de la mère
- Poids de la mère
- Statut tabagisme 0 = Non, 1 = Oui FUMÉE pendant la grossesse
- Antécédents de travail prématuré 0 = Aucun, PTD 1 = Oui
- Antécédents d'hypertension artérielle 0 = Non, 1 = Oui HT
- Présence d'irritabilité utérine 0 = Non, 1 = Oui UI

Les données ont été subdivisées en échantillons d'apprentissage (bebe_train) qui représente 349 individus et l'échantillon test [bebe_test] avec 101 individus.

1. Chargez le fichier « bebe_train.txt» dans la matrice **Dtrain** à l'aide de la procédure **csvRead()** ;

La première ligne est constituée des noms de variables et les colonnes sont séparées par le caractère tabulation

2. Affichez les 5 premières lignes des données l'aide de la commande **disp().**
3. Pour l'instant toutes les valeurs sont considérées comme des chaînes de caractères de les deux base de données .

Créez une fonction recodage qui renvoie les valeurs des descripteurs dans une matrice, et celles de la cible dans un vecteur.

Nb : pour indiquer à Scilab que les descripteurs sont en réalité numériques, il faut utiliser la fonction **evstr(),** et que la cible (7ème colonne) est une variable qualitative binaire à 2 modalités («yes»et «no») que nous codons en 1 et 0.

L'entête de la fonction : function [**descripteurs**, **cible**]=recodage(**D)**

Descripteurs: représente la matrice des variables prédictives :

cible : représente le vecteur associé à la variable cible.

Source des données : http://www.agrocampus-ouest.fr/math/livreR/ozone.txt

4. Appliquez la fonction sur l'échantillon d'apprentissage (**Dtrain)**.

https://help.scilab.org/docs/5.4.1/fr_FR/functions.html

Nb : pour avoir la distribution de fréquence des classe. vous utilisez la fonction **tabul()**.

5. Construisez un modèle prédictif (modelLD2 : un objet produit par la procédure **nan_train_sc() *,** en utilisant la Librairie « Nan ».

Quelques indications :

La procédure **nan_train_sc ()** de la librairie Nan [Il faut avoir installé et chargé la Toolbox «Nan» pour pouvoir poursuivre.] permet de réaliser une analyse discriminante ;

« **.** »Pour accéder aux champs d'un objet par exemple : **disp(**modelLD2**.weight**) pour avoir les coefficients de modèle d'analyse discriminante.

Evaluation du modèle sur un échantillon test :

Nb : Pour les questions (1) et (2), il s'agit de reproduire les commandes utilisées pour l'échantillon d'apprentissage, mais sur le l'échantillon test.

1. Chargez le fichier « bebe_test.txt» dans la matrice **Dtrain** à l'aide de la procédure **csvRead()** ;
2. Codez l'échantillon test en respectant le schéma utilisé pour le fichier d'apprentissage
3. Créez une fonction qui permet d'élabore la matrice de confusion et le taux d'erreur global, à partir des valeurs observées de **bebe_test** et des prédictions.

L'entête de la fonction :function[**MC**, **ERR_RATE**]=test_classifier(**classifier**, **descripteurs**, **cible**)

Nb : Elle prend en entrée le classifieur à évaluer, la matrice des descripteurs, le vecteur des valeurs de la variable cible. Elle renvoie en sortie la matrice de confusion et le taux d'erreur.

4. Appliquez la fonction sur l'échantillon test pour élaborer la matrice de confusion et le taux d'erreur.
5. Interpréter le taux d'erreur .

Source des données : http://www.agrocampus-ouest.fr/math/livreR/ozone.txt