POLYTECHNIQUE MONTRÉAL

MTH8302
Modèles de régression et d'analyse de la variance
Devoir 2
distribution : 31 mai 2018
remise : 12 juin 2018 - 23h59 (le plus tard)

Ce travail est réalisé individuellement par chaque étudiant inscrit au cours.
Chaque étudiant le fait SEUL sans demander de l'aide à d'autres.
En apposant sa signature ci-dessous, l'étudiant (e) certifie sur son honneur avoir fait ce travail SEUL.
L'obtention des résultats présentés et la rédaction de ce travail ne fait l'objet d'aucun plagiat, partiel ou total.

Information concernant le plagiat à Polytechnique : http://www.polymtl.ca/etudes/ppp/index.php

Exigences pour la rédaction du rapport **consulter la page 4 du plan de cours**

http://www.groupes.polymtl.ca/mth6301/mth8302/Autres/2018-MTH8302-PlanCours.pdf

Compléter l'information suivante et transmettez cette page comme la page 1 de votre rapport de devoir.

MTH8302 Modèles de régression et d'analyse de variance
NOM _____ **PRÉNOM** _____

MATRICULE _____ **SIGNATURE** _____

- **Transmettre votre rapport par courriel à** bernard.clement@polymtl.ca
- **Nom suggéré pour le fichier à transmettre : NomFamille-matricule-MTH8302-Devoir2.pdf**

TABLEAU CORRECTION

	valeur	obtenu
No 5-BostonHousing	30	
No 6-BodyFat-Femme	30	
No 7-Penta	30	
Qualité	10	
TOTAL	100	

- **Les données pour la réalisation du devoir sont disponibles sur le site WEB du cours**

http://www.groupes.polymtl.ca/mth6301/MTH8302.htm/

No 5 Étude de modélisation avec plusieurs méthodes

Données = BostonHousing.sta

Description des données

Harrison, D, Rubenfeld, D. (1978)
Hedonic (House) Prices and the Demand for Clean Air
J. of Environmental Economic and Management, v.5, pp. 81-102
Combined information from 10 separate governmental and education sources
506 census tracts (CT) in city of Boston of the year 1970

But étude de la relation entre 11 indicateurs de la qualité de vie et la valeur d'une résidence

Le fichier de 506 observations est divisé en 2 groupes

GROUP = M pour le développement des Modèles (405 observations: 80% des observations)
les données M sont surlignées (vert) et constituent un filtre;
la modélisation statistique est basée sur ces données (405 obs.)
voir *Tools...selections conditions...edit*
GROUP = T pour Tester le modèle (101 observations: 20% des observations)
pour inclure ce groupe dans une analyse, éditer le filtre

INDICATEURS

X1	CRIM	: CRIMe Rate Per Capita by town
X2	NOX	: Nitric OXide concentration (parts per 10 million)
X3	AGE	: Proportion of owner occupied units built prior to 1940
X4	DIS	: Weighted DIStances to five Boston employment centers
X5	RM	: Average number of RooMs per dwelling
X6	LSTAT	: % of the Lower STATus of the population
X7	RAD	: Index of accessibility to RADical highways
X8	CHAS	: CHASrles river dummy variable (1 if census tract bounds the river; 0 otherwise)
X9	INDUS	: Proportion of non-retail INDUStrial business acres per town
X10	TAX	: Full value property TAX rate per $10,000
X11	PT	: Pupil-Teacher ratio by town
	RLZ	: Proportion of Residential Land Zoned for lots over 25,000 sq.ft.

disponible dans le fichier mais elle ne sera pas employée

RÉPONSE

Y	MV	: Median Value of owner occupied-homes (in $1000's)

Les modèles seront développés uniquement avec le groupe M des 406 observations.

QUESTIONS

5a) Ajustez un *Modèle de Régression Ordinaire* (MRO) de Y basé sur les 11 variables X1,..., X11

5b) Les données présentent-elles un problème de multi colinéarité?

5c) Développez un *Modèle de Régression avec Sélection pas à pas Avant* (*Forward Stepwise*) (MRF)

5d) Développez un *Modèle de Régression avec Sélection pas à pas Arrière* (*Backward Stepwise*) (MRB)

Complétez le tableau 5d de la page suivante qui résume les modèles.

5e) Comparez les prédictions des 3 modèles sur l'ensemble test T constitué des 101 observations.
Choisir le meilleur modèle selon des critères; préciser la nature de ces critères.

Tableau 5d : synthèse des modèles

Var	Nom	coefficient	MRO ordinaire	MRF sélection avant	MRB sélection arrière
X0	GENÉRAL intercepte	b0			
X1	CRIM	b1			
X2	NOX	b2			
X3	AGE	b3			
X4	DIS	b4			
X5	RM	b5			
X6	LSTAT	b6			
X7	RAD	b7			
X8	CHAS	b8			
X9	INDUS	b9			
X10	TAX	b10			
X11	PT	b11			
		SS resid résiduelle			
		MSE = sigma^2 (ANOVA)			
		R^2			
		$R^2_{ajusté}$			

remarque : **laisser la cellule vide si la variable n'est pas retenue dans le modèle**

No 6 Étude d'un modèle de régression multiple problématique

Données = BodyFat-F.sta

Données physiologiques de 20 femmes en santé, âgées entre 25 et 35 ans

La mesure de l'indice de gras (BodyFat) est compliquée, longue et couteuse:
on doit faire l'immersion de la personne dans l'eau.
Peut-on développer un modèle fiable qui permettrait de prédire plus simplement et plus rapidement Y_BodyFat avec les variables faciles à mesurer :

- **X1_epTricep : épaisseur peau triceps**
- **X2_circHanches : circonférence hanches**
- **X3_circBras : circonférence du milieu du bras**

QUESTIONS

On pense qu'il est naturel que toutes les variables sont positivement corrélées entre elles et, en particulier, avec l'indice de gras.

6a) Calculez la matrice de corrélation.
Produire un scattergramme global entre toutes les variables.
La corrélation positive entre les variables est-elle valide ?

6b) Développez le modèle de régression multiple ordinaire (MRO) entre Y et X1, X2, X3.

Examinez le signe des coefficients dans le modèle MRO.
Le modèle semble-t-il satisfaisant ?
Quel est la cause ?
Que peut-on faire pour obtenir un modèle plus satisfaisant ?
Proposez 2 autres modèles, Mod1, Mod2 incorporant toutes les variables X.
pour l'obtention d'un modèle plus satisfaisante.

6c) Développez un modèle alternatif Mod1 ; précicez la méthode employée.

6c) Développez un deuxième modèle alternatif Mod2 ; précisez la méthode employée.

6d) Comparez et faites un choix entre Mod1 et Mod2 à l'aide de critères appropriés.

No 7 Étude de prédiction d'activité biologique : modélisation PLS

Données = Penta.sta

INTRODUCTION

Les nouveaux médicaments sont développés avec des produits chimiques qui sont biologiquement actifs (génie du vivant). Tester des molécules pour déceler l'activité biologique est un processus couteux et il serait utile de prédire l'activité biologique avec des mesures dont le coût serait plus faible. Il est même possible, sans même faire le composé, de calculer certaines caractéristiques comme la taille, la lidrophobicité (habileté à se dissoudre), et la polarité de groupes chimiques clés sur différents sites de la molécule ainsi que l'activité du composé. Ce domaine de recherche est appelé chimie computationnelle.

Le fichier de données, Penta, contient 31 observations et les variables

- **NOM : nom du composé**
- **15 mesures X : S1, L1,…, P5**
- **Réponse Y_logRAI : logarithme de l'activité bradykinine (enzyme de conversion)**
- **CLASSE ; classement des données : entrainement, test**

Le fichier est divisé en 2 parties; les 15 premières observations forment l'ensemble d'entrainement du modèle PLS (étude 1978 de Ufkes); les autres constituent l'ensemble test et proviennent de l'étude 1982. Les peptides utilisés dans la deuxième étude étaient différents de ceux de la première étude et la bradykinine employée dans les deux études provenait de sources différentes.

Références

Ufkes, J. G. R. et al (1978) Structure-Activity Relationships of Bradykinin-Potentiating Peptides ***European Journal of Pharmacology, vol 50, p. 119***

Ufkes, J. G. R. et al (1982) Further Studies on Structure-Activity Relationships of Bradykinin-Potentiating Peptides, ***European Journal of Pharmacology, vol 79, p. 155***

Objectif

Développer un modèle PLS basé sur la première étude et examiner sa performance à prédire les données de la deuxième étude.

QUESTIONS

7a) Développez un premier modèle PLS (noté M1) sur les seules données de test (15 premières observations) pour l'activité bradikinine. Considérez un modèle avec toutes les composantes.

7b) Développez un deuxième modèle PLS (noté M2) basé sur les 2 premières composantes seulement. Justifiez l'abandon des composantes au-delà des 2 premières.

7c) Développez un troisième modèle PLS (noté M3) basé sur les 2 premières composantes basés seulement sur les régresseurs S1 P1 S3 P3 L3 S4 L4 P4. Justifiez l'abandon des autres variables L1 S2 L2 P2 S5 L5 P5.

7d) Employez le modèle M3 pour prédire l'activité brakinine pour les données de la deuxième étude. Commentez le résultat, proposez une conclusion et, possiblement, une explication.