POLYTECHNIQUE MONTRÉAL

MTH8302
Modèles de régression et d'analyse de la variance
Devoir 3
distribution : 13 juin 2018
remise: 23 juin 2018 (au plus tard 23h55)

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.
Une copie de cette page est disponible sur le site du cours.

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-Devoir3.pdf**

TABLEAU CORRECTION

	valeur	obtenu	
No 8 - BostonHousing		30	
No 9 - Amphétamine		30	
No 10 - Assurances		30	
Qualité générale		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 8 Étude de modélisation avec MARS et réseaux de neurones

Données = BostonHousing.sta voir 2018-MTH8302-Devoir-data.stw
Ce numéro constitue une suite du numéro 5 du devoir 2.

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 avec le groupe M des 406 observations.

QUESTIONS

8a) Ajuster un Modèle de Régression MARS (MARS) de Y basé sur les 11 variables X1,…, X11 et l'ensemble M de 406 observations. Inclure des termes d'interaction d'ordre 2 dans le modèle. Inclure des graphiques qui identifient les nœuds employés dans le modèle.

8b) Développer des réseaux de neurones pour Y. Retenir les 2 meilleurs.

8c) Comparer la performance des modèles développés en 8a) et 8b) sur l'ensemble T Préciser vos critères.

8d) Compléter la comparaison en 8c) en incluant le meilleur modèle retenu lors du No5e) Présenter vos résultats dans un tableau.

8e) Identifier les forces et faiblesses des différentes méthodes de modélisation employées dans la question 5e) du devoir 2 et les questions 8a) et 8b) du devoir 3. Présenter les forces et faiblesses dans un tableau.

8f) Proposer un conclusion générale sur le processus de modélisation statistique à l'aide de modèles de régression incluant la méthode MARS et les réseaux de neuronnes.

No 9 Modèles d'analyse de la variance

Données = Amphétamine.sta voir 2018-MTH8302-Devoirs-data.stw

TITRE : effet du médicament de l'amphétamine sur le comportement

24 souris de laboratoire mâles, type albino, de poids approximativement égaux et de même souche furent utilisées dans une expérience concernant l'effet de l'amphétamine sur leur comportement lorsque privées d'eau. C'est l'objectif principal de cette expérience.
L'expérience fut réalisée en 2 parties : étude 1 et étude 2.

ÉTUDE 1

12 souris (s01, ..., s12) furent entrainées à activer un levier pour obtenir de l'eau jusqu'à ce qu'elles obtiennent un taux relativement stable d'activation. Sur la base de ce résultat, les souris furent classées en 3 catégories (lente, moyenne, vite) de vitesse initiale. Ce facteur est représenté par la variable XB_vitesse dans le fichier.
Les souris reçurent de l'amphétamique selon 4 niveaux de dosage (mg/kg). Ce facteur est représenté par la variable XC_dose dans le fichier et les modalités fixées dans cette expérience sont fixées: 0 (solution saline) / 0,5 / 1,0 / 1,8 (mg/kg). Les 4 niveaux furent administrés au hasard pour chaque souris. Une heure après réception, une séance expérimentale commence. La souris reçoit de l'eau après 2 coups (appui) sur le levier. C'est le facteur XA_levier dont la modalité est 2 dans l'étude 1. Chaque dose fut administrée 2 fois représenté par la variable rep_dose (1 et 2). Cette variable est un facteur de répétition. La réponse mesurée Y est définie par

Y = nombre de coups sur le levier / temps écoulé durant la session
temps est mesuré en seconde.

ÉTUDE 2

12 souris additionnelles (s13, ..., s24) furent utilisées pour la deuxième expérience. L'étude2 est semblable à l'étude1 sauf que les souris reçoivent de l'eau après 5 coups.
Dans ce cas la variable XA_levier = 5

ANALYSES à faire

étude1 seulement : Y en fonction de XB_vitesse, XC_dose
étude1 et étude2 combinées : Y en fonction de XA_levier, XB_vitesse, XC_dose
remarque : les modèles proposés pour l'analyse seront composés d'effets principaux et des effets d'interactions doubles

DONNÉES (extrait des 5 premières observations) le fichier global contient 192 observations

	1 ID	2 Étude	3 souris	4 XA_levier	5 XB_vitesse	6 XC_dose	7 rep_dose	8 Y
1	1	étude1	s01	2	lente	0,0	1	0,81
2	2	étude1	s01	2	lente	0,5	1	0,80
3	3	étude1	s01	2	lente	1,0	1	0,82
4	4	étude1	s01	2	lente	1,8	1	0,50
5	5	étude1	s02	2	lente	0,0	1	0,77

QUESTIONS

9a) Pour chacune des 2 études, décrire la nature et le rôle des facteurs dans le modèle d'analyse de variance/covariance qui sera employé pour faire l'analyse.

9b) Pourquoi les souris furent-elles initialement classées en catégories de vitesse?

9c) Pourquoi les souris recoivent-elles la dose d'amphétamine dans un ordre dicté par le hasard?

9d) Proposer le modèle qui sera employé pour faire l'analyse de l'étude 1.
Exécuter cette l'analyse et présenter les principaux résultats sous forme de tableaux et de graphiques. Proposer une conclusion principale de cette étude.

9e) Répondez aux mêmes questions que 9d) pour les 2 études combinées.

No 10 Modélisation statistique

Données = Assurances.sta voir 2018-MTH8302-Devoirs-data.stw

Extrait du fichier : 5 premières observations

réclamations de 788 assurés souffrant de maladie coronarienne

Les variables ont une signification évidente.
La variable de réponse est v12 = Y_coûtTotal (en $) des réclamations par l'assuré.
Les variables v4 et v6 furent recodées pour leur donner un statut de variables catégoriques.

v4_recod est un recodage de v4 (nombre inter&proc) selon le tableau suivant

v4	0	1	2	3-4	5-6-7	8 et plus
v4_recod	0	1	2	3&4	5à7	8+

v6_recod est un recodage de v6 (nb autres maladies) selon le tableau suivant

v6	0 ou 1	2 ou plus
v6_recod	0&1	2&plus

/

1 ID	2 age	3 genre	4 nombre interventions& procédures	5 v4_recod	6 nombre autres maladies	7 v6_recod	8 nombre médicaments prescrits	9 nombre visites unités soins intensifs	10 nombre complications	11 durée traitement (jr)	12 Y_coûtTotal
1	63	F	2	2	3	2&plus	1	4	0	300	179,1
2	59	F	2	2	0	0&1	0	6	0	120	319,0
3	62	F	17	8+	5	2&plus	0	2	0	353	9310,7
4	60	H	9	8+	2	2&plus	0	7	0	332	280,9
5	55	F	5	5à7	0	0&1	2	7	0	18	18727,1

Le fichier contient l'hisrorique de 788 réclamations faites à une compagnie d'asssurance par des patients soufrant de maladie coronarienne.

Le fichier est caractérisé par 12 variables dont l'interprétation est évidente.

Les variables v4 et v6 furent recodées en v3-recod et v6_recod selon le tableau dans l'entête du tableau. Ces 2 variables ont alors un statut de variable catétogorique.

QUESTIONS

10a) L'analyse statistique des données peut se faire selon plusieurs modèles.
Proposer 4 modèles statistiques que l'on peut considérer pour faire l'analyse. Présenter vos modèles en complétant le tableau

Modèle	Nom statistique (*)	Définir le rôle de chacune des variables impliquées
M1		
M2		
M3		
M4		

nom statistique (*) : consulter la page 37 du document :
http://www.groupes.polymtl.ca/mth6301/mth8302/NotesCours/2017-MTH8302-ANOVA-partie1.pdf

rôle des variables : continu, catégorique (facteur), covariable, autre

10b) Effectuer l'analyse de 2 modèles : présenter les principaux résultats seulement; résumer les conslusions de chaque modèle.

10c) Comparer les résultats de 2 modèles. Y – a-t-il des différences d'interprétations?

10d) La variable de réponse aurait t-elle dû être transformée? Justifier.