POLYTECHNIQUE MONTRÉAL

MTH8302

Modèles de régression et d'analyse de la variance

Devoir 1

distribution : 14 mai 2018

remise : 27 mai 2018 à 23h59 (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.
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**_____

Transmettez votre rapport par courriel à **bernard.clement@polymtl.ca**

Nom suggéré pour le fichier à transmettre : **aaaa_mmm_2018_MTH8302_devoirN.pdf**

aaaa = nom de famille **mmm = matricule** **N = numéro du devoir (1, 2, 3, 4)**

TABLEAU CORRECTION

	valeur	obtenu
No 1-Anscombe	30	
No 2-Vaccins	30	
No 3-Croissance	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 1 Analyse diagnostique / graphique dans les modèles statistiques

Données = Anscombe.sta

Le fichier contient 4 couples de variables, (X1, Y1), (X2, Y2), (X3, Y3) et (X4, Y4).
On considère un modèle de régression linéaire simple pour prédire Y en fonction de X pour chacun des 4 couples (X, Y):

modèle 1 : $Y1 = \beta_0 + \beta_1 X1 + \varepsilon$
modèle 2 : $Y2 = \beta_0 + \beta_1 X2 + \varepsilon$
modèle 3 : $Y3 = \beta_0 + \beta_1 X3 + \varepsilon$
modèle 4 : $Y4 = \beta_0 + \beta_1 X4 + \varepsilon$

QUESTIONS

1a) Complétez les valeurs manquantes du tableau ci bas.

modèle couple	β_0	β_1	R^2	SSreg	SSresid	SStot
1 (X1, Y1)						
2 (X2, Y2)						
3 (X3, Y3)						
4 (X4, Y4)						

R^2 : coefficient de détermination = fraction de la variation de Y expliqué par X

SSreg : somme de carrés de régression (expliquée) par le modèle

SSresid : somme de carrés résiduelle (erreur)

SStot : somme des carrés totale

Commentez le tableau.

1b) Tracez, pour chacun des 4 couples de variables, un nuage de points (« 2D scatterplots ») illustrant la variation de Y en fonction de X.
Commentez les graphiques. Faites un lien avec le commentaire en 1a).

1c) Pour chacun des 4 couples de variables, tracer un graphique des résidus en fonction de la variable explicative X. Commentez. Faites un lien avec le commentaire en 1b).

1d) Pour chacun des 4 couples de variables, tracer un graphique des résidus sur échelle de probabilité gaussienne. Commentez.

1e) Considérons le couple (X3, Y3). La droite de régression est-elle affectée s'il s'avère que l'observation (X3 =13 Y3 = 12,74) est le résultat d'une erreur et peut être éliminée.
Refaire les calculs sans cette observation.
Le modèle est-il adéquat avec cette observation?
Que devient alors la valeur de R^2 sans cette observation?

1f) Proposez une conclusion générale pour ce numéro.

No 2 Régression logistique – programme de sensibilisation

Données = Vaccins.sta

Étude sur l'efficacité d'un programme de sensibilisation vaccin contre la grippe.
Un Centre Local de Santé Communautaire (CLSC) a envoyé un dépliant publicitaire encourageant les personnes âgées à recevoir un vaccin contre la grippe. Une étude subséquente, un échantillon de 159 personnes furent choisies au hasard et on leur demanda s'il avait reçu le vaccin.
La variable de réponse Y est

reçu le vaccin **Y_reçuVaccin = 1 = oui**

pas reçu le vaccin **Y_reçuVaccin = 0 = non**

Le fichier contient trois autres variables potentiellement explicatives.

X1_âge : âge de la personne

X1_catAge : âge représenté par 5 valeurs typiques : 52 - 57 - 62 - 67 - 72

52 = 54 et moins | 57 = 55 à 59 | 62 = 60 à 64 | 67 = 65 à 69 | 72 = 70 et plus

X2_indSanté : indice de sensibilisation à sa santé - échelle 0 à 100

0 = aucune sensibilité **100 = sensibilité très élevée**

X3_genre : sexe de la personne : **F = Femme** **H = homme**

colonnes ajoutées 9-10-11-12
tableau croisé = colonne 7 X colonne 3
Y0 = nombre pas vaccinés Y1 = nombre vaccinés
selon les catégories d'âge (colonne 3) pour l'ensemble de tous les répondants

QUESTIONS

2a) **Ajusté un modèle de régression logistique entre Y et X1_âge.**

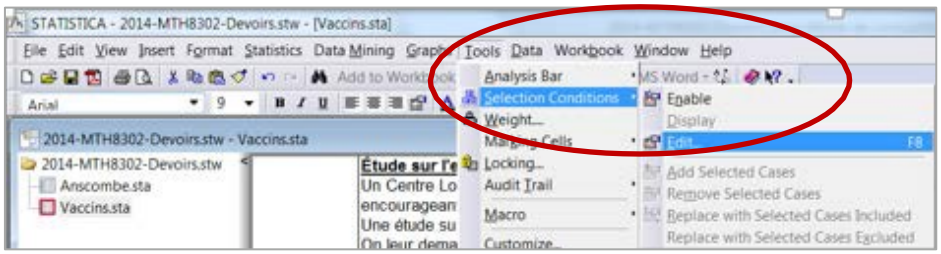
Les variables X2 et X3 ne sont pas tenues en compte dans le modèle.

Module de Statistica ***Nonlinear Estimation ….Quick logit regression***

2b) **Refaire 2a) pour les hommes seulement.**

Recommandation : imposer un filtre sur les données avec

Tools…Selection conditions…edit



2c) **Refaire 2b) pour les femmes seulement.**

2d) **Comparez les 3 modèles de 2a) 2b) 2c) : effet de l'âge et du sexe sur Y**

2e) **Refaire l'analyse 2a) avec le tableau résumé des colonnes 9 / 10 / 11.**
Comparer les modèles 2a) et 2e).

No 3 Régression non linéaire

Données = croissance.sta

L'observation des phénomènes de croissance de matière biologique, animale, végétale,.. donne souvent lieu à une courbe en forme de S (sigmoide). La fonction de croissance mesurée est représentée par une fonction non linéaire car le phénomène est souvent caractérisé par une évolution lente au début suivi par une croissance rapide et se terminant par une stabilisation progressive. Plusieurs fonctions ont été proposées pour modéliser le phénomène. Ces fonctions sont paramétrées par des paramètres a , b , c, d, …Les fonctions suivantes sont parmi les plus employées pour modéliser les phénomènes de croissance.

Voir ModelesNonLineaire.pdf http://www.groupes.polymtl.ca/mth6301/MTH8302.htm

Gompertz	$Y = a*\exp[-\exp(b - cx)]$
Logistique 3P	$Y = a / [1 + \exp(b - cx)]$
remarque :	**ne pas confondre avec le modèle logistique obtenu avec a = 1**
Morgan-Mercer	$Y = (b*c + a\,x^d) / (c + x^d)$
Weibull	$Y = a - b \exp(-c\,x^d)$
Probit	$Y = a\,\Phi(b + c\,x)$

Φ (u) : fonction de répartition de la distribution normale centrée-réduite

---

Le fichier de données proposé pour ce numéro est un exemple typique de données de croissance. Employez le module *Nonlinear Estimation* de *STATISTICA* pour réaliser cet exercice.

QUESTIONS

3a) **Tracer le graphique des données.**
3b) **Ajuster la fonction de Gompertz.**
3c) **Ajuster la fonction Logistique.**
3d) **Ajuster la fonction Weibull.**
3e) **Proposer la meilleure fonction choix de fonction pour modéliser les données. Préciser le critère employé pour faire le choix.**

Information sur la fonction Weibull

La distribution de Weibull a pour fonction de densité f avec des paramètres b, c et q positifs :

$$f(x ; b, c, q) = c / b*[(x - q) / b]^{(c-1)}\, e^{\{-[(x - q) / b]^{**c}\}} \quad \text{pour } 0 < q <= x < \infty \quad b > 0 \quad c > 0$$

b	**est le paramètre d'échelle de la distribution**
c	**est le paramètre de forme de la distribution**
q	**est le paramètre de position de la distribution**
e	**= 2,71828…constante d'Euler**

La fonction de répartition F de la distribution de Weibull est:

(1) $F(x) = e^{\{-[(x - q) / b]^{**c}\}}$

On peut modifier cette fonction en ajoutant un facteur de décalage d et un facteur multiplicatif a:

(2) $G(x) = d + a*e^{\{-[(x - q) / b]^{**c}\}}$

G(x) n'est plus une fonction de répartition sauf si d = 0 et a = 1
En particulier, si on pose q = 0, on obtient:

(3) $H = d + a*e^{\{-[(x) / b]^{**c}\}}$

Avec un changement de notation, on obtient l'équation (4).

(4) $Y = a - b \exp(-c\,x^{**d})$

Statistica, avec l'estimation de Quasi-Newton, présente une meilleure convergence avec la forme (3) qu'avec la forme (4) si on spécifie les contraintes (b > 0, c > 0). Il est possible de le faire avec la forme (4) mais il faut ajuster les contraintes.