Thierry Duchesne thierry.duchesne@mat.ulaval.ca

Travail pratique 1

À remettre **AVANT** le vendredi 16 novembre 2018, 23 h 59.

Peut être fait en équipes de 1, 2 ou 3 individus (équipes de 2 fortement recommandées).

Instructions de remise du devoir

- **IMPORTANT :** Inscrivez votre équipe sur le site web du cours avant le vendredi 26 octobre 2018, 23 h 59.
- Remettez un seul devoir par équipe, sous la forme de deux fichiers :
  (i) un fichier **PDF nommé** `Dev1-NomFamille1_NomFamille2.pdf` qui contient les solutions aux exercices théoriques ainsi que les rapport demandés (avec leurs annexes respectives) pour les questions pratiques (**les solutions manuscrites numérisées ne seront PAS acceptées**) ;
  (ii) un fichier **texte nommé** `Dev1-NomFamille1_Nomfamille2.txt` qui contient tout le code informatique utilisé pour les questions pratiques.

IMPORTANT : La qualité du français et de la présentation sera évaluée et comptera pour 10% de la note totale.

Partie théorique

T-1 Exercices 1, 10 et 12 des séries d'exercices suggérés du chapitre 1.

T-2 Exercices 2, 4 (seulement les parties (a) et (b)) et 5 des séries d'exercices suggérés du chapitre 2.

T-3 Exercice 1 de la série d'exercices suggérés du chapitre 3.

Partie pratique

Pour chaque question de la partie pratique, produisez un rapport dans lequel vous expliquez et justifiez clairement toutes les étapes ayant mené à votre modèle final. (N'oubliez pas d'effectuer un diagnostic de multicollinéarité avant de commencer à construire les modèles !) Ensuite répondez aux questions spécifiques qui vous sont posées.

P-1 Weisberg (2005) présente les données d'une étude menée à l'université Berkeley dans les années 1920 où 28 enfants ont été suivis de l'âge de 2 ans jusqu'à l'âge de 18 ans. Les données de l'étude sont contenues dans le fichier `weisberg56.dat`, disponible sur le site web du cours dans la sous-section où vous avez trouvé le présent document. La définition de chacune des variables du jeu de données vous est fournie en annexe du présent devoir. Votre objectif est de trouver un modèle de régression qui explique la valeur du somatotype à l'âge de 18 ans (variable `SOMA`) à l'aide des variables mesurées aux âges de 2 ans et de 9 ans.

(a) Interprétez la relation entre les variables mesurées à l'âge de 2 ans et le somatotype selon votre modèle final.

(b) Un enfant de 9 ans ayant les mesures ci-dessous se présente à vous. Utilisez votre modèle final pour donner une prévision ponctuelle et un intervalle de confiance à 95% pour son somatotype à l'âge de 18 ans.

`WT2=13  HT2=90  WT9=41  HT9=141  LG9=31.5  ST9=73`

P-2 Le jeu de données `processed.cleveland.data` est un fichier texte avec colonnes séparées par des virgules qui donne la valeur de 13 variables explicatives (voir description en annexe à ce devoir). L'objectif de cet exercice est de construire un modèle de régression qui estime la probabilité de diagnostic de maladie coronarienne positif (donc définissez une variable réponse Y qui vaut 1 si la variable `num` prend une valeur supérieure à 0 et qui vaut 0 sinon) à partir de la valeur des 13 variables explicatives.

Selon votre modèle final, quels sont les facteurs qui semblent associés à une hausse du risque d'un diagnostic positif de maladie coronarienne ?

P-3 Le jeu de données `ausprivauto0405` est disponible dans le package R `CASdatasets` (package qui n'est pas sur CRAN, mais allez chercher le fichier `CASdatasets_1.0-6.tar.gz` sur le site web officiel du package). Il contient la valeur de 3 caractéristiques du véhicule, 2 caractéristiques de l'assuré, la proportion de l'année pendant laquelle l'assuré a été couvert par sa compagnie d'assurance auto ainsi que s'il y a eu réclamation (oui ou non), le nombre de réclamations et le montant total des réclamations pendant la période couverte (plus de détails sur les variables en annexe). L'objectif est de construire un modèle afin de voir s'il y a une association entre les caractéristiques du véhicule et du client et le nombre de réclamations (variable réponse). Décrivez précisément votre modèle final. La valeur relative du véhicule est-elle associée au nombre espéré de réclamations ? Si oui, interprétez cet effet.

Attention, lors de la construction de votre modèle, n'oubliez pas de considérer le besoin de tenir compte de la surdispersion ou d'une variable d'offset.

Annexe

Variables du fichier `weisberg56.dat`

- `ID` : numéro d'identification de l'enfant
- `WT2` : masse (en kg) à l'âge de 2 ans
- `HT2` : grandeur (en cm) à l'âge de 2 ans
- `WT9` : masse (en kg) à l'âge de 9 ans
- `HT9` : grandeur (en cm) à l'âge de 9 ans
- `LG9` : circonférence de la jambe (en cm) à l'âge de 9 ans
- `ST9` : force de préhension (en kg) à l'âge de 9 ans
- `WT18` : masse (en kg) à l'âge de 18 ans
- `HT18` : grandeur (en cm) à l'âge de 18 ans
- `LG18` : circonférence de la jambe (en cm) à l'âge de 18 ans
- `ST18` : force de préhension (en kg) à l'âge de 18 ans
- `SOMA` : somatotype, un indice d'obésité sur une échelle continue, qui prend de faibles valeurs (1 ou moins) pour les gens très maigres et une forte valeur (7 ou plus) pour les gens très obèses

Variables du fichier `processed.cleveland.data`

Les variables du fichier sont entrées dans le même ordre que celui qui suit.

- `age` : âge en années
- `sex` : 1 = homme, 0 = femme
- `cp` : nature des douleurs à la poitrine, variable **qualitative à 4 modalités**, où 1 dénote l'angine typique, 2 l'angine atypique, 3 une douleur non anginienne et 4 une douleur asymptomatique
- `trestbps` : tension artérielle au repos (en mm Hg) à l'admission à l'hôpital
- `chol` : cholestérol sanguin en mg/dl
- `fbs` : indicatrice qui vaut 1 si le taux de sucre sanguin à jeun > 120 mg/dl et qui vaut 0 sinon
- `restecg` : résultat de l'électrocardiogramme au repos, variable **qualitative à 3 modalités**, où 0 signifie normal, 1 signifie anomalie des ondes ST-T et 2 signifie hypertrophie probable du ventricule gauche
- `thalach` : pouls maximum atteint
- `exang` : indicatrice indiquant la présence d'angine induite par l'exercice (1 pour oui, 0 pour non)
- `oldpeak` : baisse dans ST induite par l'exercise par rapport au repos
- `slope` : pente du segment de ST lors de l'exercice maximal, variable **qualitative à 3 modalités** soit 1 pour ascendante, 2 pour plate et 3 pour descendante
- `ca` : nombre de vaissaux sanguins majeurs colorés par fluroscopie
- `thal` : variable **qualitative à 3 modalités** où 3 = normal, 6 = défaut réparé, 7 = défaut réparable
- `num` : la variable réponse que nous cherchons à prédire est $Y = 1$ si `num`> 0 et $Y = 0$ si `num`$= 0$

Variables du jeu de données `CASdatasets_1.0-6.tar.gz` du package `CASdatasets`

- `Exposure` : proportion de l'année pendant laquelle l'assuré(e) est couvert(e)
- `VehValue` : valeur relative du véhicule (mesure continue)
- `VehAge` : âge du véhicule sous la forme de variable **qualitative à 4 modalités**
- `VehBody` : type de véhicule sous la forme de variable **qualitative à 13 modalités**
- `Gender` : sexe de l'assuré(e) sous la forme de variable **qualitative à 2 modalités**
- `DrivAge` : âge de l'assuré(e) sous la forme de variable **qualitative à 6 modalités**
- `ClaimOcc` : PAS UTILISÉE DANS CE DEVOIR
- `ClaimNb` : nombre de réclamations, **variable réponse**
- `ClaimAmount` : PAS UTILISÉE DANS CE DEVOIR