DM Mise à Niveau R

Votre devoir est à me renvoyer par mail avant le 18 septembre à l'adresse anna.bonnet@upmc.fr, aux formats htlm et Rmd. Votre fichier Rmd doit pouvoir être compilé lorsque le jeu de données **icecream** se trouve dans le même répertoire et que les packages nécessaires ont été installés (vous n'utiliserez que les packages étudiés en cours ainsi que **ggplot2movies** pour les données).

Exercice 1

La loi de Benford est une loi semi-empirique qui affirme que dans les nombres que l'on rencontre dans la vie courante, le premier chiffre significatif n'est pas équiréparti dans l'ensemble $\{1, \dots, 9\}$, mais qu'il est distribué selon la loi suivante :

$$p_i = \mathbb{P}(\text{premier chiffre significatif égal à i}) = \log_{10}(\frac{i+1}{i})$$

où log_{10} est le logarithme décimal, défini par $\log_{10}(x) = \frac{\log(x)}{\log(10)}$.

Au début des années 1990, un économiste américain a remarqué que lorsque les gens falsifiaient les comptes d'une société, ils avaient tendance à utiliser trop de 5 et de 6 comme premiers chiffres significatifs, c'est-à-dire plus que n'en prévoit la loi de Benford, qui par ailleurs est bien vérifiée pour ce genre de données numériques.

1. Expliciter et représenter avec un diagramme en bâtons la loi de Benford.
2. Ecrire une fonction qui prend en entrée un nombre et qui renvoie son premier chiffre significatif.
3. Charger les données **movies** qui se trouvent dans le package **ggplot2movies**. Représenter la distribution empirique suivi par le premier chiffre significatif de la variable **budget** (en enlevant tous les NA et les 0). Commenter.
4. On souhaite tester si la distribution du premier chiffre significatif de ces données suit bien la loi de Benford. Quel test allons-nous réaliser ? Le mettre en oeuvre et conclure.

Exercice 2

A partir du jeu de données **icecream**, nous allons étudier la consommation de glace aux Etats-Unis sur une période de 30 semaines du 18 Mars 1950 au 11 Juillet 1953. Les variables sont la consommation (Consumption en pintes par habitant), le salaire hebdomadaire (Income en dollars), le prix des glaces (Price en dollars), la température (Temp en degré fahrenheit) et la catégorie socio-professionnelle (sc).

1. Charger le jeu de données **icecream** avec le nom des colonnes en utilisant la fonction **read.table** et en spécifiant correctement les arguments sep, row.names et header. Précisez la nature des variables (qualitative ou quantitative) et faites une analyse rapide des données.

2. Créer un jeu de données comprenant seulement les variables quantitatives. Découper aléatoirement ce jeu de données en deux échantillons: un échantillon d'apprentissage (avec 70% des données) et un échantillon de test.

3. Sur le jeu de données d'apprentissage, représenter les nuages de points de **cons** en fonction de(s) variable(s) quantitative(s). Effectuer ensuite la régression linéaire de la variable **cons** en fonction de toute(s) le(s) variable(s) quantitative(s). Ce modèle sera appelé **modele1**. Afficher un résumé et interpréter le résultat (variables significatives, ...).

4. Effectuer une procédure de sélection de variable par le critère d'information bayésien (BIC). Quelles sont les variables sélectionnées ?

5. Effectuer une nouvelle régression linéaire avec uniquement les variables retenues en question précédente, ce modèle sera appelé **modele2**. Refaire ensuite la procédure de sélection de variables.

6. Dans un vecteur, stocker les valeurs de **cons** prédites par le modèle **modele1** pour chaque individu de l'échantillon de test. Construire de même un vecteur à partir de **modele2**. Utiliser pour cela la fonction **predict**.

7. Notons Y la variable **cons**. Pour $i = 1, ..., N_t$ où N_t est le nombre d'observation de l'échantillon de test, on note $\hat{Y}_i^j$ la prévision par le modèle j du i-ème individu de l'échantillon test, et Y_i la valeur de Y observée sur le i-ème individu de l'échantillon test. Calculer

$$\text{EQM1} = \frac{1}{N_t}\sum_{i=1}^{N_t}(\hat{Y}_i^1 - Y_i)^2 \text{ et } \text{EQM2} = \frac{1}{N_t}\sum_{i=1}^{N_t}(\hat{Y}_i^2 - Y_i)^2.$$

Interpréter.

9. Représenter **cons** en fonction de(s) variable(s) qualitative(s). Effectuer une analyse de la variance. Conclure sur l'effet de(s) variable(s) qualitative(s) sur la consommation de glace.

Exercice 3

On va travailler à nouveau sur les données **movies** du package **ggplot2movies**.

1. Tracer la note des films en fonction de l'année de sortie. Colorer les points selon que le film dure plus ou moins de 180 minutes.

2. Créer un sous-ensemble qui contient les colonnes title, year, length, budget, rating et uniquement les films de moins de 300 minutes. Arranger le tableau de sorte qu'il soit ordonné par longueur de film décroissante.

3. Réarranger le jeu de données afin qu'il contienne une seule colonne "Genre" (un film qui est à la fois classé comme Comédie et Action aura donc deux lignes, l'une dans laquelle son genre est Comédie, l'autre dans laquelle c'est Action).

4. Refaire le graphique de la question 1 avec un graphique pour chaque genre.

5. Trouver le budget moyen et la note moyenne pour chaque genre, en ignorant les valeurs NA.

6. Créer une variable Satisfaction avec trois niveaux : Mauvais pour les notes inférieures à 5, Moyen pour les notes entre 5 et 7.5, Bon pour les notes supérieures à 7.5.

7. Créer une variable avec deux niveaux : Long pour les films de plus de 3h et Court pour les autres. Représenter avec des boxplots les notes obtenues dans ces deux groupes. Faire un test pour savoir s'il y a une différence significative de notes entre les deux groupes.