Modélisation statistique : TP 1

Nous allons étudier les données "Vulnerability" (Patt et al., PNAS - 2009). La question est la suivante : les pays les moins développés sont-ils plus vulnérables aux changements climatiques ? Les auteurs ont voulu expliquer `ln_death_risk`, log du risque mortel dû aux évènement climatiques en fonction

- du log du nombre d'évènements climatiques `ln_death_risk`
- du log de la fertilité `ln_fertility`
- de l'indice de développement humain `hdi` (United Nations)
- du log de la population `ln_pop`

Ils concluent que le développement socio-économique a un lien sur la fragilité aux événements climatiques, et ce lien pourrait se révéler dans le deuxième quart du 21ième siècle.

1. Avant de commencer,
   (a) lancer Rstudio
   (b) récupérer le fichier `notebook_TP1.Rmd` et les données "vulnerabilty" sur ma page web.
   (c) créer un répertoire dans vos documents pour ce TP, y mettre tous les fichiers associés. N'oubliez pas de conserver tous vos fichiers.
   (d) charger et installer le package `tidyverse`
   (e) changer l'option du chunck pour que le code précédent n'apparaissent pas sur la preview.
2. Chargement des données et visualisation
   (a) Charger les données dans R. Vérifier le type de chaque variable.
   (b) Faire un scatterplot. Repérer graphiquement (prendre des notes)
      i. les variables linéairement liées à ln_death_risk,
      ii. les éventuelles variables à transformer (lien non-linéaire)
      iii. les éventuelles corrélations linéaires entre variables explicatives.
3. Modèle linéaire simple
   (a) Faire un premier modèle linéaire `fit_univ` avec seulement la variable ln_events. Etudier le `summary`
   (b) Que contient l'objet `fit_univ` ?
   (c) Que vaut $\hat{\beta}$?
   (d) Représenter graphiquement la droite estimée et les données brutes. Que pensez vous de ce modèle ?
   (e) Peut-on accepter le test de $\mathcal{H}_0 : \beta_1 = 0$? Qu'est ce que ça signifie ?

(f) Que valent le R^2 et le R^2 ajusté ?

NB : On définit le R^2 par

$$0 \leq R^2 = \frac{\|X\hat{\beta} - \bar{Y}\mathbf{1}\|^2}{\|Y - \bar{Y}\mathbf{1}\|^2} = 1 - \frac{\|Y - X\hat{\beta}\|^2}{\|Y - \bar{Y}\mathbf{1}\|^2} \leq 1$$

et le R^2 ajusté du nombre de paramètres par

$$R^2_{Adj} = 1 - \frac{(n-1)(1-R^2)}{(n-p-1)} \leq 1$$

Attention à la dimension $p+1$: c'est le nombre de variables explicatives $p+1$ pour le coefficient constant (associé à $(1, \ldots, 1)$).

(g) Pour un nouvel individu, on a observé ln_events = 3.4. Quelle est votre prédiction pour son ln_death_risk (fonction `predict`) ? Quel est l'intervalle de confiance pour votre prédiction ?

4. Modèle linéaire multiple

   (a) Faire un second modèle linéaire avec tous les variables explicatives, faire l'analyse du `summary`.

   (b) Comparer le R^2 et le R^2 ajusté à ceux du premier modèle.

   (c) Via un test de Fisher comparer ce nouveau modèle au précédent. Lequel préférez-vous ?

   (d) Essayer d'ajouter la variable hdi^2, comparer au modèle précédent. Quel modèle préférez-vous ?

5. Sélection de modèle

   (a) Faire une sélection de modèle via l'AIC puis le BIC (fonction `step`) et interpréter le modèle final.

   (b) Comparer les R^2, R^2 ajusté et AIC des différents modèles (null, univarié, multivarié complet, multivarié après sélection).

1 Observations isolées et aberrantes, leviers et résidus

(a) Simulation dans un modèle linéaire gaussien

   i. Simuler $n = 20$ observations suivant le modèle linéaire gaussien

$$Y_i = 2 + 4X_i + \epsilon_i$$

   avec $X_i = i$ et $\epsilon_i \sim \mathcal{N}(0, \sigma^2 = 2)$. Construire un jeu de données `data` avec ces deux variables.

   ii. Représenter les points (X_i, Y_i). Estimer un modèle linéaire pour la régression de Y sur X

   iii. Faire les diagnostics pour rechercher des observations influentes ou aberrantes.

(b) La 20ième observation du modèle précédent devient maintenant **isolée** : $X_{20} = 30$.

   i. Simuler Y_{20} dans le modèle précédent, i.e. en posant $Y_{20} = 2 + 4X_i + \epsilon_{20}$.

ii. Représenter les points (X_i, Y_i).

iii. Faire les diagnostics pour rechercher des observations influentes ou aberrantes.

(c) La 20ième observation du modèle précédent devient maintenant **aberrante** : $X_{20} = 20$ mais $Y_{20} = 2 + 4X_i + \epsilon_{20} - 10$.

i. Simuler Y_{20} comme indiqué.

ii. Représenter les points (X_i, Y_i).

iii. Faire les diagnostics pour rechercher des observations influentes ou aberrantes.

(d) La 20ième observation du modèle précédent devient maintenant à la fois **isolée et aberrante.**

i. Simuler Y_{20} comme indiqué.

ii. Représenter les points (X_i, Y_i).

iii. Faire les diagnostics pour rechercher des observations influentes ou aberrantes.

2 Jeu de données Vulnerability : diagnostics sur les observations et les variables

Nous allons analyser continuer notre analyse du jeu de données "Vulnerability" (Patt et al., PNAS - 2009).

Repartir du modèle avec toutes les variables

```
fit = lm( ln_death_risk ~ ln_urb + ln_events + ln_fert + hdi + ln_pop , data = vul)
```

(a) Faire, pour chaque variable explicative, un graphique pour vérifier le lien liénaire avec `ln_death_risk`.

(b) Faire les diagnostics de corrélation entre variables explicatives. Enlever des variables pour obtenir une matrice X bien conditionnée.

(c) Faire une recherche d'individus aberrants et influents (attention à ne pas enlever trop d'individus !!).

(d) Faire les diagnostics sur les résidus.