Modélisation statistique : TP 2

Avant de commencer,

- récuperer le fichier R_TP2.Rmd et les données "vulnerability" sur ma page web
- créer un répertoire dans vos documents pour ce TP, y mettre tous les fichiers associés.

1 Observations isolées et aberrantes, leviers et résidus

1. Simulation dans un modèle linéaire gaussien
   (a) 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.
   (b) Représenter les points (X_i, Y_i). Estimer un modèle linéaire pour la régression de Y sur X
   (c) Faire les diagnostics pour rechercher des observations influentes ou aberrantes.
2. La 20ième observation du modèle précédent devient maintenant **isolée** : $X_{20} = 30$.
   (a) Simuler Y_{20} dans le modèle précédent, i.e. en posant $Y_{20} = 2 + 4X_i + \epsilon_{20}$.
   (b) Représenter les points (X_i, Y_i).
   (c) Faire les diagnostics pour rechercher des observations influentes ou aberrantes.
3. La 20ième observation du modèle précédent devient maintenant **aberrante** : $X_{20} = 20$ mais $Y_{20} = 2 + 4X_i + \epsilon_{20} - 10$.
   (a) Simuler Y_{20} comme indiqué.
   (b) Représenter les points (X_i, Y_i).
   (c) Faire les diagnostics pour rechercher des observations influentes ou aberrantes.
4. La 20ième observation du modèle précédent devient maintenant à la fois **isolée et aberrante.**
   (a) Simuler Y_{20} comme indiqué.
   (b) Représenter les points (X_i, Y_i).
   (c) 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)
```

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

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

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

4. Faire les diagnostics sur les résidus.