Régression robuste aux valeurs extrêmes

Ce travail doit être remis avant le **18 février à 17h** sur Moodle.

Données

Cet exercice est basé sur le jeu de données *gapminder* du package du même nom.

> Jennifer Bryan (2017). gapminder: Data from Gapminder. R package version 0.3.0. https://CRAN.R-project.org/package=gapminder

Ce jeu de données inclut l'espérance de vie (*lifeExp*), la population (*pop)* et le PIB par habitant (*gdpPercap*) pour 142 pays et 12 années (aux 5 ans entre 1952 et 2007).

```
library(gapminder)
str(gapminder)
```

```
## tibble [1,704 x 6] (S3: tbl_df/tbl/data.frame)
##  $ country  : Factor w/ 142 levels "Afghanistan",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ continent: Factor w/ 5 levels "Africa","Americas",..: 3 3 3 3 3 3 3 3 3 3 ...
##  $ year     : int [1:1704] 1952 1957 1962 1967 1972 1977 1982 1987 1992 1997 ...
##  $ lifeExp  : num [1:1704] 28.8 30.3 32 34 36.1 ...
##  $ pop      : int [1:1704] 8425333 9240934 10267083 11537966 13079460 14880372 12881816 13867957 163
##  $ gdpPercap: num [1:1704] 779 821 853 836 740 ...
```

1. Effet du PIB et du temps sur l'espérance de vie

a) Visualisez d'abord les données d'espérance de vie en fonction du PIB par habitant et de l'année. Il est suggéré de représenter le logarithme de *gdpPercap* et de séparer les différentes années, par exemple avec des facettes dans *ggplot2*: `... + facet_wrap(~year)`.

Quelles tendances générales observez-vous? Semble-t-il y avoir des valeurs extrêmes qui pourraient influencer fortement un modèle de régression? Si oui, essayez d'identifier ces données dans le tableau en vous basant sur la position des points dans le graphique.

b) Réalisez une régression linéaire (`lm`) pour déterminer l'effet du PIB par habitant, de l'année et de leur interaction sur l'espérance de vie. Pour faciliter l'inteprétation des coefficients, effectuez les transformations suivantes sur les prédicteurs:

- Prenez le logarithme de *gdpPercap* et normalisez-le avec la fonction `scale`. *Rappel*: `scale(x)` soustrait de chaque valeur de `x` leur moyenne et divise par leur écart-type, donc la variable résultante a une moyenne de 0 et un écart-type de 1; elle représente le nombre d'écarts-types au-dessus ou en-dessous de la moyenne.
- Remplacez *year* par le nombre d'années écoulées depuis 1952.

Interprétez la signification de chacun des coefficients du modèle, puis consultez les graphiques de diagnostic. Les suppositions du modèle linéaire sont-elles respectées?

c) Comparez le résultat du modèle en (b) avec deux alternatives plus robustes aux valeurs extrêmes: la régression robuste basée sur le bipoids de Tukey (fonction `lmrob` du package *robustbase*) et la régression

de la médiane (fonction `rq` du package *quantreg*, en choisissant seulement le quantile médian). Expliquez comment les estimés des coefficients et leurs erreurs-types diffèrent entre les trois méthodes.

Note: Utilisez l'option `showAlgo = FALSE` en appliquant la fonction `summary` au résultat de `lmrob`, pour simplifier le sommaire.

d) Superposez les droites de régression des trois modèles sur le graphique en (a). Avec `ggplot`, vous pouvez utiliser la fonction `geom_smooth` avec `method = "lm"` pour la régression linéaire et `method = "lmrob"` pour la régression robuste. Pour la régression de la médiane, vous pouvez utiliser `geom_quantile` tel que vu dans les notes.

2. Variation des effets par quantile

a) D'après votre observation des données en 1(a), serait-il utile de modéliser différents quantiles de l'espérance de vie en fonction des prédicteurs? Justifiez votre réponse.

b) Réalisez une régression quantile avec les mêmes prédicteurs qu'en 1(b), avec les quantiles suivants: `(0.1, 0.25, 0.5, 0.75, 0.9)`. Utilisez la fonction `plot` sur le sommaire de la régression quantile et décrivez comment l'effet des prédicteurs varie entre les quantiles.

c) Superposez les droites de régression quantile au graphique des données. Les tendances pour chaque quantile semblent-elles affectées par des valeurs extrêmes?

Note sur les comparaisons internationales

Bien que ce jeu de données soit utile pour illustrer les concepts de régression robuste et de régression quantile, soulignons que ce type d'analyse statistique comparant des variables mesurées au niveau national comporte plusieurs limites:

- On ne peut pas supposer que les associations détectées s'appliquent à une échelle plus petite (ex.: le lien entre espérance de vie et revenu en comparant les moyennes nationales n'est pas nécessairement le même que le lien entre espérance de vie et revenu au niveau des individus habitant chaque pays).
- Les moyennes calculées dans différents pays ne sont pas des observations indépendantes, car les conditions environnementales, sociales et économiques sont corrélées entre pays proches.
- Il y a de nombreux facteurs qui différencient les pays, donc il est difficile d'interpréter une association comme un lien de cause à effet.

Beaucoup d'articles, en particulier dans les domaines des sciences sociales, ont été publiés au sujet des méthodes à suivre pour réaliser ce type de comparaisons internationales (*cross-country comparisons*).