

Exo1 : régression linéaire simple

Claude Grasland & Gbètoton Nadège Djossou

07/06/2020

Objectif

On se propose dans ce TD de modéliser la relation entre PIB par habitant (X) et émission de CO2 des pays africains (Y) en 2018 à l'aide d'une relation linéaire de type $Y = f(X)$. On commencera par utiliser un modèle de régression linéaire simple en soulignant les multiples violations des hypothèses qu'il entraîne. Puis on proposera deux solutions alternatives, l'une en retirant les valeurs exceptionnelles, l'autre en transformant les variables X et Y de façon logarithmique.

1. PREPARATION DES DONNEES

1.1 Importation des données

```
don<-read.csv2("data/afrika_don.csv")
```

1.2 Sélection des variables

On décide de renommer les deux variables choisies X et Y

- X : PIB en \$/habitant
- Y : CO2 en tonnes/habitant

```
don$X<-don$PIB  
don$Y<-don$CO2HAB
```

1.4 Extraction du tableau à analyser

On ne garde que les colonnes iso3, name, reg, X et Y. Et on élimine les lignes comportant des valeurs manquantes à l'aide de la fonction *complete.case()*

```
tab<-don[,c("iso3", "name", "X", "Y")]  
tab<-tab[complete.cases(tab), ]
```

1.5 Astuce : stockage des textes d'habillage

On prépare un ensemble de textes que l'on pourra utiliser pour l'habillage de nos graphiques. Cela évitera de devoir ensuite les retaper à chaque fois.

```
nomX <- "PIB ($/hab)"  
nomY <- "Pollution (t. de CO2/hab)."  
titre <- "Les pays Africains en 2018"  
note <- "Source : Rapport sur le développement humain 2020"
```

2. ANALYSE DES VARIABLES X et Y

2.1 La distribution de X

Calculer les paramètres principaux et commentez les

- **Commentaire :**

Faire un histogramme

- Histogramme rapide
- Histogramme amélioré
- **Commentaire :**

Tester la normalité

- **Commentaire :**

Examiner la présence de valeurs exceptionnelles

- **Commentaire :**

2.2 La distribution de Y

Calculer les paramètres principaux

- **Commentaire :**

Faire un histogramme

- Histogramme rapide
- Histogramme amélioré
- **Commentaire :**

Tester la normalité

- **Commentaire :**

Examiner la présence de valeurs exceptionnelles

- **Commentaire :**

3. CORRELATION

3.1 Visualiser la relation entre X et Y

- Graphique rapide
- Graphique amélioré
- **Commentaire : :**

3.2 Tester la significativité de la relation entre X et Y

Coefficient de Pearson

- **Commentaire :**

Coefficient de Spearman

- Commentaire :

4. REGRESSION LINEAIRE

4.1 Calculer l'équation de la droite $Y = aX+B$

- Commentaire :

4.2 Visualiser la droite

- Commentaire:

4.3 Calculer les valeurs estimées et les résidus

- Commentaire :

4.4 Sauvegarder les résultats du modèle

5. DIAGNOSTICS

Avant de tirer des conclusions hâtives sur les résidus, il est préférable de vérifier si les hypothèses fondamentales du modèle de régression ont bien été respectées. On va utiliser pour cela quatre graphiques de bases fournis par R et des tests présents dans le package `car` (acronyme de “Companion for Applied Regression”).

5.1 Autocorrélation des résidus

- Commentaire :

5.2 Normalité des résidus

- Commentaire :

5.3 Homogénéité des résidus

- Commentaire :

5.4 Absence de valeurs exceptionnellement influentes

- Commentaire :

5.5 Tous les tests d'un coup

Une fois que l'on a bien compris les tests précédents, on peut afficher les quatre graphiques correspondant en une seule commande :

6. AUTRES MODELES

Sans reprendre en détail toutes les étapes de l'analyse, proposez deux variantes du modèle initial, l'une en retirant les valeurs exceptionnelles, l'autre en transformant les variables X et Y à l'aide d'une fonction préalablement à leur mise en relation.

6.1 Modèle linéaire sans valeurs exceptionnelles.

On décide de retirer les trois valeurs exceptionnellement influentes qui ont été repérées dans la première analyse et de refaire une régression linéaire.

Correction du tableau

Corrélation

Régression

Visualisation

Diagnostics

- **Commentaire :**

6.2 Modèles non linéaires

Il est toujours ennuyeux de retirer des valeurs exceptionnelles car on risque d'en trouver des nouvelles et c'est un processus sans fin. Il s'agit en outre d'une démarche critiquable si on effectue le retrait des valeurs sans raisons objectives. Il est donc préférable d'essayer de garder toutes les valeurs mais de chercher à transformer les variables X et Y pour construire des fonctions différentes. On utilise classiquement quatre modèles (linéaire, exponentiel, logarithmique, puissance) selon que l'on applique ou non des transformations linéaires à X et Y.

Examen visuel des quatre modèles

- **Commentaire :**

Calcul des coefficients de corrélation

- **Commentaire :**

Préparation des données

On crée un nouveau tableau de données

Régression

Visualisation

Diagnostics

- **Commentaires :**

Représenter la forme finale du modèle $Y = f(X)$

- **Commentaire :**