Rendu-TD-9

2025-11-13

Exercice 1. Quels facteurs démographiques ou économiques expliquent les émissions de CO2 des nations ?

Contexte

Dans cet exercice, on cherche à étudier la relation entre les émissions de CO2 des pays et différents facteurs démographiques et économiques susceptibles de les influencer.

Pour cela, on utilise le jeu de données `co2.per.capita`, qui rassemble des informations sur les émissions de CO2 par habitant (per capita) pour 169 pays à travers le monde.

Les données utilisées dans cet exercice sont extraites de la base **Our World in Data** (https://ourworldindata.org/) qui regroupe des informations mondiales sur les émissions de CO2 et les gaz à effet de serre. Ce jeu de données inclut également des indicateurs provenant du Programme des Nations Unies pour le développement (PNUD), relatifs à la démographie et au développement économique des différents pays.

On lit les données dans le tableau `df` de la manière suivante

```
df = read.csv2("./data/co2_per_capita.csv")
View(df)
```

Chaque ligne représente un pays. Les colonnes contiennent les variables suivantes :

1. "**country**" : Nom du pays
2. "**population**" : Taille de la population
3. "**co2.per.capita**" : Émissions annuelles de dioxyde de carbone (CO2) par habitant basées sur la production, mesurées en tonnes exprimées sur une échelle logarithmique (log10). Ce chiffre est basé sur les émissions territoriales et ne tiennent pas compte des émissions provenant des biens échangés.
4. "**gni.per.capita**" : Le revenu national brut (RNB/GNI) est la somme totale d'argent en milliards de dollars US gagnée par la population et les entreprises d'un pays divisée par le nombre d'habitants. Il est utilisé pour mesurer et suivre la richesse d'un pays d'une année sur l'autre. Ce chiffre comprend le produit intérieur brut (PIB) de la nation et les revenus qu'elle reçoit de l'étranger. Le RNB est exprimé sur une échelle logarithmique (log10).
5. "**life.expectancy**" : Espérance de vie d'un ressortissant.
6. "**hdi**" : Indice de développement humain. L'indice de développement humain (IDH/HDI) correspond à un indice calculé chaque année par le Programme des Nations unies pour le développement. L'IDH vise à évaluer le niveau de développement des pays en se fondant non pas sur des données strictement économiques, mais sur la qualité de vie de leurs ressortissants. L'indice prend en compte l'espérance de vie, le niveau d'éducation et le logarithme du revenu brut par habitant.

Question 1

Représenter, à l'aide d'un diagramme en barres, les émissions de CO2 par habitant des 50 pays présentant les valeurs les plus élevées.

Les pays devront être classés par ordre décroissant d'émissions, et leurs noms affichés en abscisse du graphique.

Solution.

Les émissions annuelles de dioxyde de carbone par habitant sont exprimées en tonnes (les données sont en échelle logarithmique log10). Compléter le code ci-dessous (commandes utiles : `barplot`,`axis`)

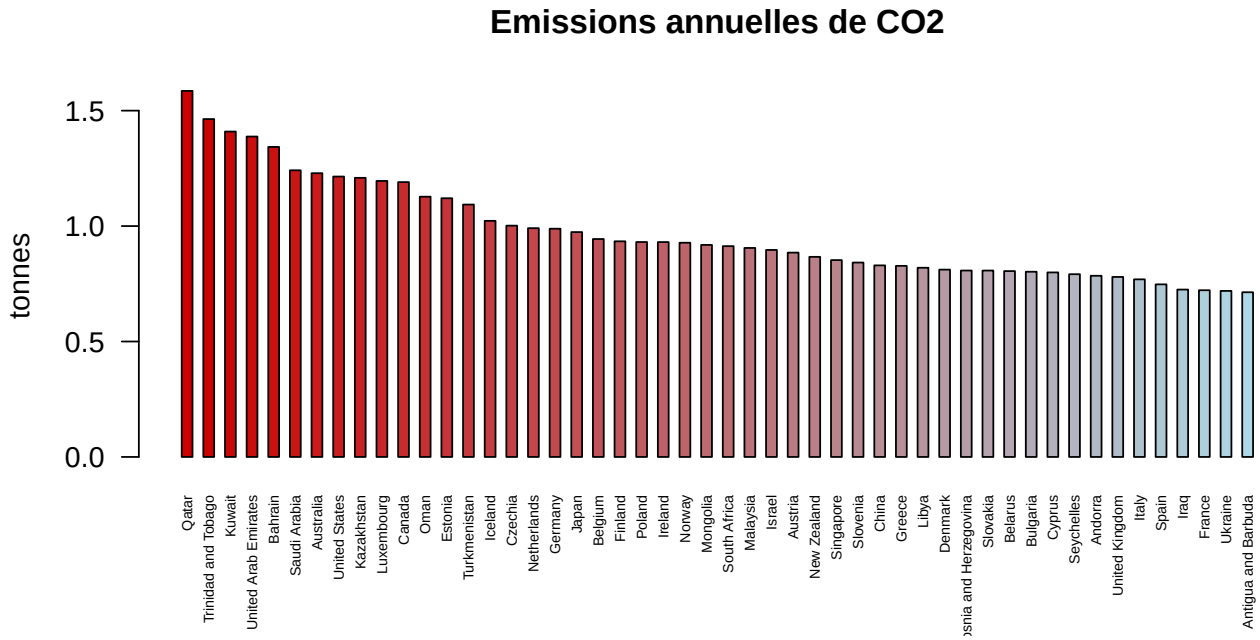
```
#  remplacer eval = FALSE par eval = TRUE
## nombre de pays top émetteurs
n_pays = 50

## palette de couleur pour représenter les émissions les plus intenses en dégradé
ma_palette <- colorRampPalette(c("red3", "lightblue"))

## trier les valeurs d'émissions des pays top émetteurs
emission_n_pays <-  sort(df$co2.per.capita, decreasing = TRUE)[1:n_pays]

## Diagramme en barre
barplot(emission_n_pays,
        col = ma_palette(n_pays),
        space = 1,
        las = 1,
        main = "Emissions annuelles de CO2",
        xlab = "",
        ylab = "tonnes")

## noms des pays sur l'axe horizontal
axis(side = 1, at = 2*(1:n_pays) - 0.5,
     labels = df$country[order(df$co2.per.capita, decreasing = TRUE)[1:n_pays]],
     tick = FALSE,
     cex.axis = .5, las = 3)
```



Question 2

Identifier les pays dont les émissions de CO2 par habitant dépassent 15 tonnes. Les 15 pays correspondants devront être présentés par ordre alphabétique.

Solution. Compléter le code ci-dessous en précisant la condition appropriée (commande utile : `with`).

```
#  remplacer eval = FALSE par eval = TRUE
## Sélection des pays répondant à la condition
with(df, country[co2.per.capita > log(15, base = 10)])
```

```
##  [1] "Australia"            "Bahrain"              "Canada"
##  [4] "Kazakhstan"           "Kuwait"               "Luxembourg"
##  [7] "Qatar"                "Saudi Arabia"         "Trinidad and Tobago"
## [10] "United Arab Emirates" "United States"
```

Question 3

Identifier les pays dont les émissions de CO2 par habitant sont inférieures à 100 kg, soit 0,1 tonne.

Solution. Inclure un code R permettant de sélectionner et d'afficher les pays correspondant à cette condition.

```
with(df, country[co2.per.capita < log(0.1, base = 10)])
```

```
## [1] "Burundi"                  "Central African Republic"
## [3] "Chad"                     "Malawi"
## [5] "Niger"                    "Rwanda"
```

Question 4

Ajuster un modèle de régression linéaire simple pour étudier la relation entre les logarithmes (log10) des émissions de CO2 par habitant (`co2.per.capita`) et du revenu national brut par habitant (`gni.per.capita`). Afficher le résumé du modèle (`summary`). À partir de ce résumé, déterminer :

1. l'équation de la droite de régression estimée,
2. la proportion de la variance des émissions de CO2 qui est expliquée par le revenu national brut.

Solution. Inclure le code R et répondre clairement aux deux points ci-dessus.

```
#Modèle de la régression linéaire simple
mod = lm(df$co2.per.capita ~ df$gni.per.capita)

#Affichage du resumé
resume = summary(mod)
resume
```

```
##
## Call:
## lm(formula = df$co2.per.capita ~ df$gni.per.capita)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.9535 -0.1809 -0.0192  0.1619  0.8079
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       -4.24254    0.16185  -26.21   <2e-16 ***
## df$gni.per.capita  1.12107    0.03959   28.32   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.263 on 167 degrees of freedom
## Multiple R-squared:  0.8277, Adjusted R-squared:  0.8266
```

```
## F-statistic:   802 on 1 and 167 DF,  p-value: < 2.2e-16
```

On note y les émissions de CO2 par habitants en log10 et x le revenu par habitant. L'équation de la droite est :
$y = -4.24 + 1.12 * x$
La proportion de variance expliquée des émissions de CO2 par le revenu national brut correspond à R^2 et est de : 0.8277

Question 5

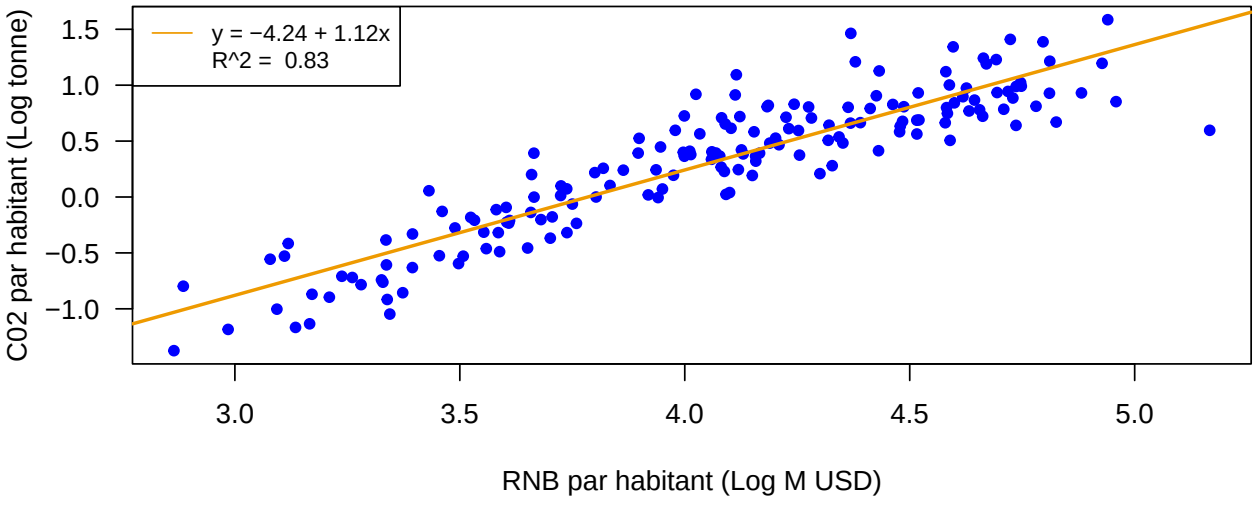
Tracer un graphique représentant les données et la droite de régression linéaire ajustée. Ajouter une légende en haut à gauche du graphique. La légende doit indiquer l'équation de la droite de régression et le coefficient de détermination du modèle.

Solution. Compléter le code R permettant de produire le graphique et la légende.

```r
#  remplacer eval = FALSE par eval = TRUE
## On trace le graphique des données : la quantité de CO2 émise par habitant en fonction
# du revenu par habitant
plot(df[c("gni.per.capita", "co2.per.capita")],
     cex = .8, col = "blue", pch = 19, las = 1,
     xlab = "RNB par habitant (Log M USD)",
     ylab = "CO2 par habitant (Log tonne)")

## On trace la droite de régression linéaire du modèle
abline(mod, col = "orange2" , lwd = 2)

## On écrit la légende
legend("topleft", col = c("orange2",NA), lty = 1, cex = .8,
       legend = c("y = -4.24 + 1.12x",
                  paste("R^2 = ",
                        round(summary(mod)$r.squared, 2))))
```



Question 6

À partir du modèle de régression ajusté :

1. Calculer l'augmentation prévue des émissions de CO2 par habitant lorsque le revenu national brut par habitant augmente de 10%.
2. Calculer l'effet sur les émissions de CO2 si le revenu national brut est multiplié par 10.

Solution. Compléter le texte suivant avec les résultats numériques. Lorsque le revenu national brut augmente de 10% les nouvelles émissions de CO2 par habitant se calculent avec : $log(y_{nv}) = -4.24 + 1.12 * log(1.1x)$.
L'augmentation prévue est donc de :

$$log(y_{nv}) - log(y) = 10^{1.12log(1.1)} = 0.113$$

soit de 11.2%

Lorsque le revenu national est multiplié par 10, les nouvelles émissions de CO2 par habitant se calculent avec : $log_{10}(y_{nv}) = -4.24 + 1.12 * log_{10}(10x)$.
Les emissions sont donc multipliées par :

$$\frac{y_{nv}}{y} = 10^{1.12} = 13.18$$

Une augmentation de 10% du revenu national brut correspond à une augmentation de 0.112 tonnes des émissions de CO2 par habitant. Si le revenu national brut est multiplié par 10, les émissions de CO2 sont multipliées par un facteur 13.18.

Question 7

L'indice de développement humain (IDH/HDI) d'un pays est un indice composite qui prend en compte l'espérance de vie, le niveau d'éducation et le logarithme du revenu brut par habitant.

1. Vérifier si l'IDH est fortement corrélé au logarithme du revenu brut par habitant.
2. Ajuster un modèle de régression linéaire pour expliquer l'IDH à partir du logarithme du revenu brut par habitant.

Interpréter l'erreur résiduelle dans ce modèle.

Solution. Pour interpréter l'erreur résiduelle, noter que le logarithme du revenu brut par habitant apparaît à la fois dans la variable explicative et dans le calcul de l'IDH. Réfléchir donc à la part de variation de l'IDH qui n'est pas expliquée par ce logarithme.

```
#On calcule le coefficient de corrélation linéaire R
r = cov(df$gni.per.capita,df$hdi)/sqrt(var(df$gni.per.capita)*var(df$hdi))
r
```

```
## [1] 0.9657294
```

```
#Modèle de régression linéaire pour expliquer IDH à partir du logarithme du revenu brut par habitant
mod_hdi = lm(df$hdi~df$gni.per.capita)
resume = summary(mod_hdi)
resume
```

```
##
## Call:
## lm(formula = df$hdi ~ df$gni.per.capita)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.133477 -0.019587  0.005584  0.027298  0.088171
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)       -0.451786   0.024608  -18.36   <2e-16 ***
## df$gni.per.capita  0.289394   0.006019   48.08   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 0.03999 on 167 degrees of freedom
## Multiple R-squared:  0.9326, Adjusted R-squared:  0.9322
## F-statistic:  2312 on 1 and 167 DF,  p-value: < 2.2e-16
```

On trouve un coefficient de corrélation R de 0.97 environ donc l'IDH est fortement corrélé au logarithme du revenu brut par habitant.
L'erreur résiduelle dans ce modèle correspond à l'effet de l'espérance de vie et du niveau d'éducation sur l'indicateur HDI.

Question 8

Réfléchir aux facteurs autres que la richesse d'un pays (RNB) qui pourraient expliquer les émissions de CO2 (le RNB est la somme totale d'argent gagnée par la population et les entreprises d'un pays). Dans le modèle qui relie le logarithme des émissions de CO2 par habitant au logarithme du revenu brut par habitant, proposer une interprétation de l'erreur résiduelle du modèle.

Solution. L'erreur résiduelle représente la partie de la variation des émissions de CO2 (échelle log10) qui n'est pas expliquée par le logarithme du RNB.

Les autres facteurs que la richesse d'un pays qui pourraient expliquer les émissions de CO2 peuvent être : -Le mix énergétique (Part des énergies fossiles, Part de l'hydraulique,...) -Les politiques publiques favorisent-elles l'écologie ? (USA en ont pas grande chose à faire, France un peu +) -La culture du consumérisme (Indépendamment de combien on produit, es-ce que on consomme beaucoup et comment) -La taille de la population (Peut être qu'il y a des économies d'échelles à faire sur les infrastructures) -La taille du pays (Besoin ou non de beaucoup de routes pour désenclaver l'ensemble du territoire)

L'erreur résiduelle reflète l'ensemble des facteurs non inclut dans le modèle comme par exemple ceux cités au dessus.

Question 9

On considère :

- les résidus des émissions de CO2 (log10) expliquées par le logarithme du RNB.
- les résidus de l'indice de développement humain (IDH) expliqué par le logarithme du RNB.

Ajustez un modèle de régression linéaire où les émissions résiduelles (log10) sont expliquées par les indices de développement résiduels.

Interpréter les résultats : ce modèle suggère-t-il que la consommation et la richesse industrielle (représentées par le RNB) sont les principaux facteurs des émissions de CO2, comparés aux efforts pour la santé ou l'éducation ? Si c'est le cas, comment cela se traduit-il en termes de significativité statistique ?

Solution. Pour répondre, définir clairement les variables utilisées (résidus des émissions de CO2 et résidus de l'IDH corrigé pour la richesse) et analyser le modèle.

```
#  remplacer eval = FALSE par eval = TRUE
residus_emission <- mod$residuals
residus_hdi <-  mod_hdi$residuals
```

et ajuster le nouveau modèle

```
#  remplacer eval = FALSE par eval = TRUE
mod_residuals <-  lm(residus_emission ~  residus_hdi)
summary(mod_residuals)
```

```
##
## Call:
## lm(formula = residus_emission ~ residus_hdi)
```

```
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.88063 -0.17383 -0.03284  0.16396  0.80964
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 5.914e-18  2.013e-02   0.000    1.000
## residus_hdi 6.718e-01  5.063e-01   1.327    0.186
##
## Residual standard error: 0.2616 on 167 degrees of freedom
## Multiple R-squared:  0.01043,    Adjusted R-squared:  0.004508
## F-statistic: 1.761 on 1 and 167 DF,  p-value: 0.1863
```

On effectue un modèle pour regarder si les emisions de CO2 qui ne dépendent pas du RNB dependent des efforts pour la santé et l'éducation (HDI sans le RNB) On trouve une valeur de $R^2 = 0.01$ donc on peut dire que les emissions de CO2 ne dépendent pas des efforts pour la santé et l'éducation.

Question 10

Observer les résultats obtenus avec le modèle suivant à deux prédicteurs :

```
summary(lm(co2.per.capita ~ gni.per.capita + hdi , data = df))
```

```
##
## Call:
## lm(formula = co2.per.capita ~ gni.per.capita + hdi, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.88063 -0.17383 -0.03284  0.16396  0.80964
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -3.9390     0.2806 -14.040  < 2e-16 ***
## gni.per.capita   0.9267     0.1522   6.089 7.62e-09 ***
## hdi              0.6718     0.5078   1.323    0.188
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2624 on 166 degrees of freedom
## Multiple R-squared:  0.8295, Adjusted R-squared:  0.8274
## F-statistic: 403.7 on 2 and 166 DF,  p-value: < 2.2e-16
```

Identifiez les points remarquables concernant les effets estimés pour chaque prédicteur et leur significativité statistique.

Solution. Interpréter le résumé et commenter les observations pertinentes.

La lecture des coefficients du modèle indique que le CO2 par tête peut être modélisé comme suit:

$$log_{10}CO2 = -3.93 + 0.93 \times log_{10}GNI + 0.67 \times HDI$$

On a une forte significativité statistique pour b_0 et b_1 : entre 0 et 0.001 et une significativité statistique faible pour b_2 : 1. Pour le RNB, on observe que $\hat{b}_1 = 0.9267$ avec une p_valeur de test inférieur à 0.05 donc on rejette l'hypothèse $b_1 = 0$ Pour HDI, on observe que $\hat{b}_2 = 0.6718$ avec une p_valeur de test supérieur à 0.05

donc on ne peut pas rejetter l'hypothèse $b_2 = 0$. Si $b_2 = 0$, cela signifierait que, quand on prend déjà en compte le GNI dans la régression, on n'a pas besoin d'ajouter l'HDI.

De plus, R^2 est pratiquement identique à celui trouvé dans la régression entre les emissions de CO2 et le revnu national brut par habitant ce qui confirme bien la faible importance de HDI dans la quantité des emissions de CO2.

De plus, dans le modèle du CO2 expliqué uniquement par le GNI, le coefficient associé au GNI était de 1.12. Il est ici de 0.93, ce qui semble cohérent avec le fait que l'un des paramètres pris en compte dans le calcul de l'HDI est le GNI. Une part de l'explication des émissions de CO2 par le GNI se retrouve donc dans la part d'explication de l'IDH, bien qu'elle ne soit pas significative.