

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

```
summary(tab$X)
```

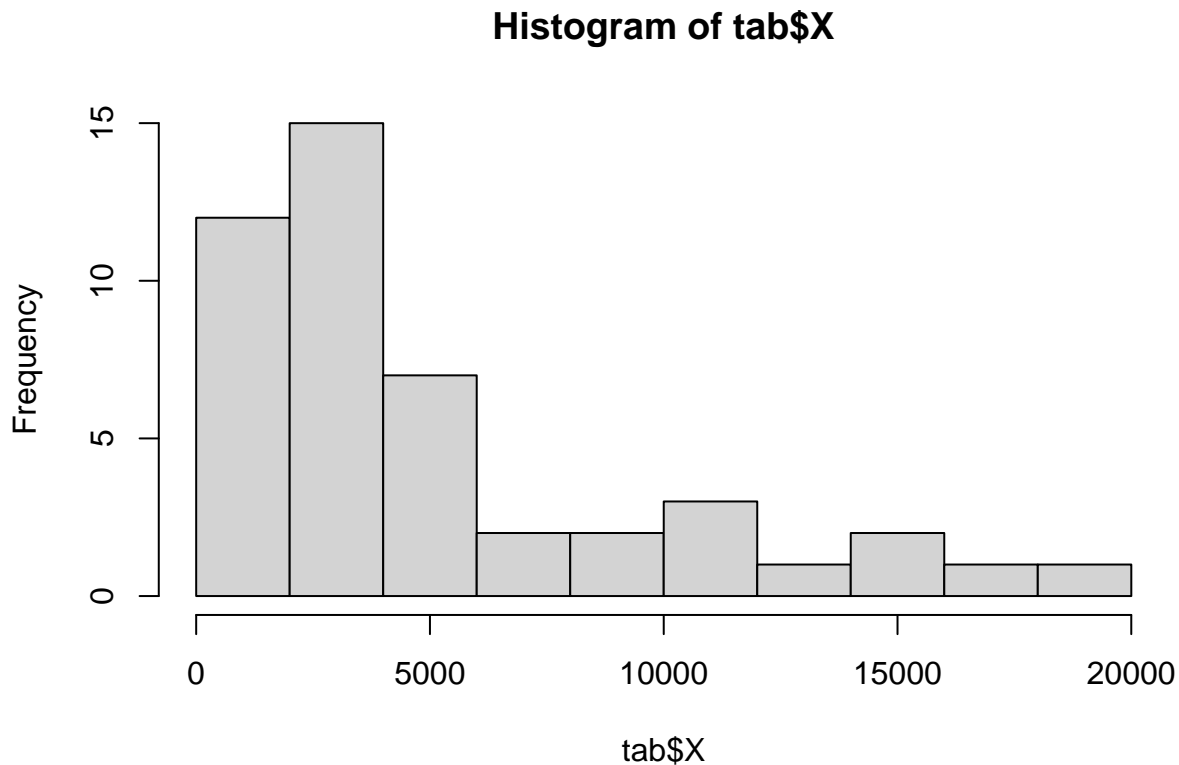
```
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  756.6   2014.9   3289.4   5168.8   6437.0  19458.9
```

- **Commentaire :** Le PIB par habitant des pays africains varie entre 756 et 19459. Il est en moyenne de 5169. La moitié des pays ont un taux compris entre Q1 (2015) et Q3 (6437)

Faire un histogramme

- Histogramme rapide

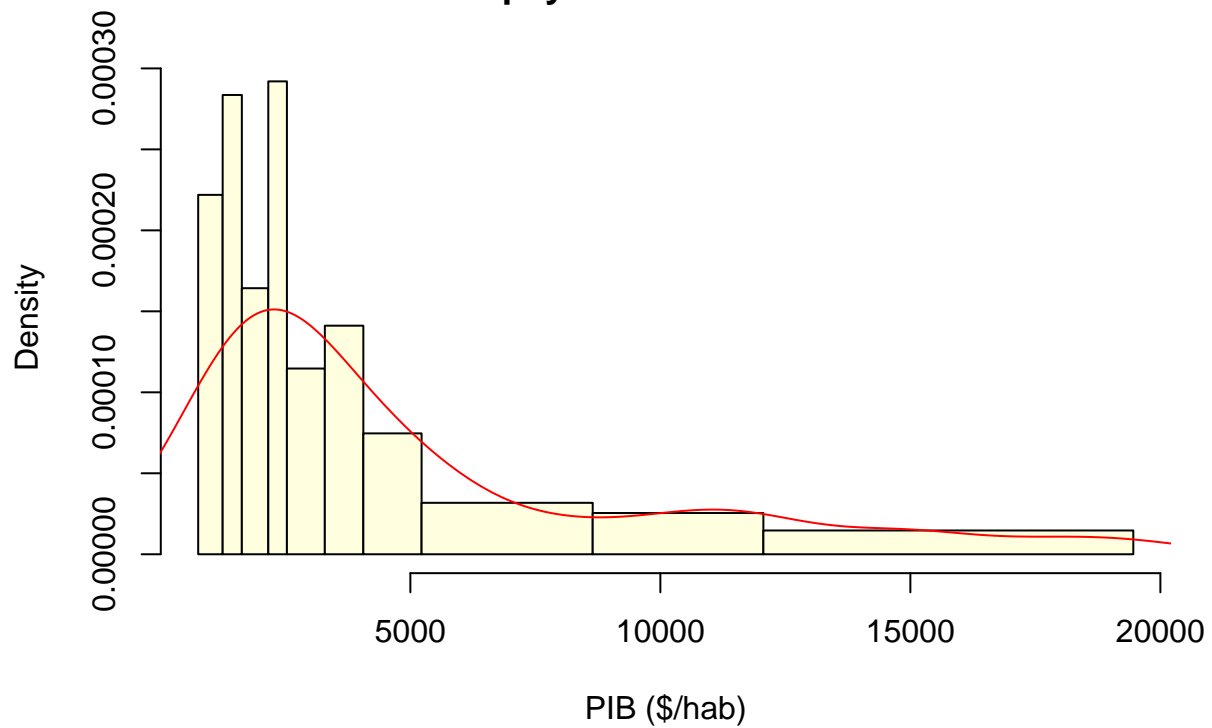
```
hist(tab$X)
```



- Histogramme amélioré

```
hist(tab$X,
      xlab=nomX,
      breaks=quantile(tab$X, c(0,0.1,0.2,0.3,0.4,0.5,0.6,0.7,0.8,0.9,1)),
      main = titre,
      sub = note,
      col = "lightyellow")
lines(density(tab$X),col="red")
```

Les pays Africains en 2018

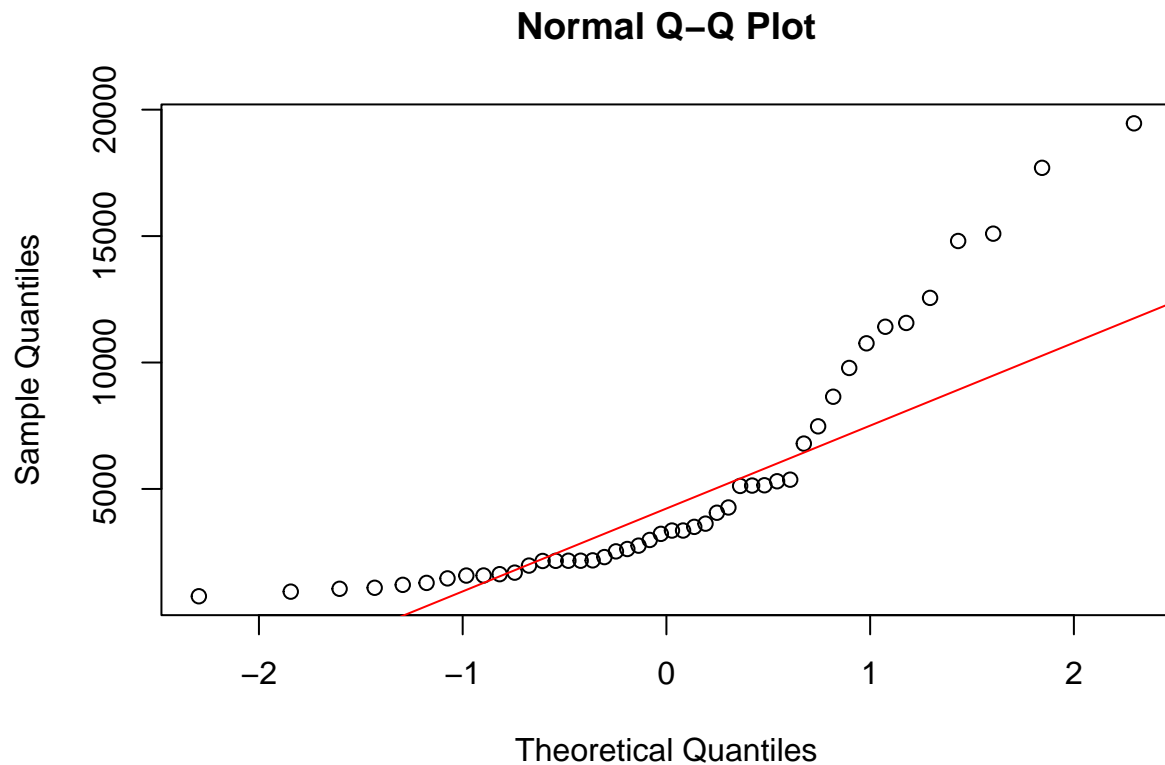


Source : Rapport sur le développement humain 2020

- **Commentaire :** La distribution semble unimodale mais fortement asymétrique à gauche.

Tester la normalité

```
# Graphique
qqnorm(tab$X)
qqline(tab$X, col = "red")
```



```
# test
shapiro.test(tab$X)
```

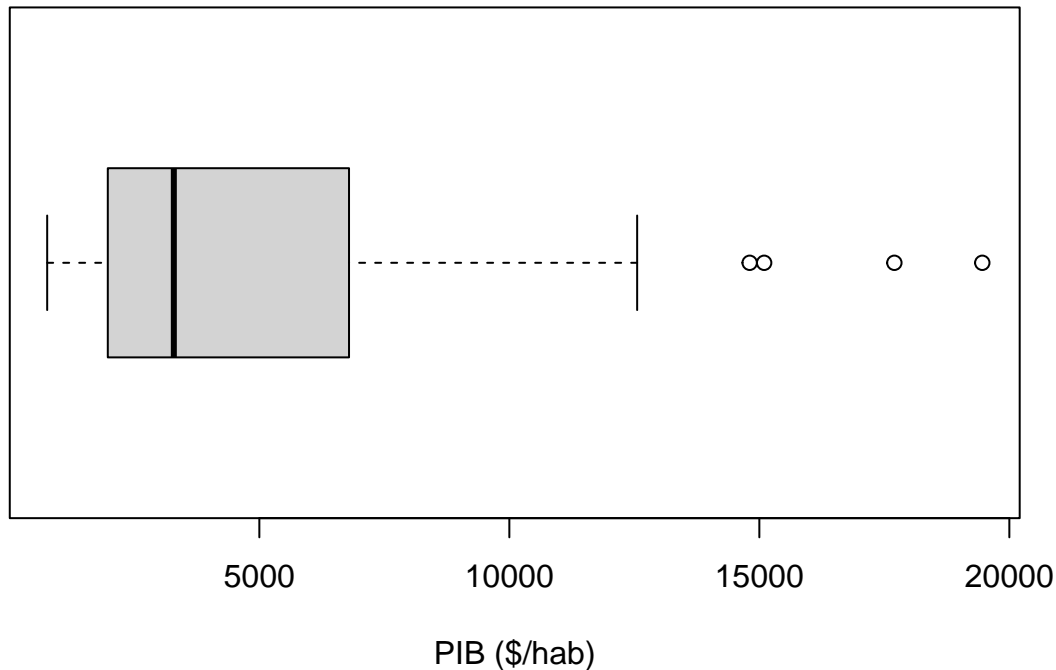
```
##
##  Shapiro-Wilk normality test
##
## data:  tab$X
## W = 0.796, p-value = 1.584e-06
```

- **Commentaire :** Le graphique montre que la distribution ne suit pas une loi gaussienne, ce qui est confirmé par le test de Shapiro-Wilks ($p < 0.001$)

Examiner la présence de valeurs exceptionnelles

```
boxplot(tab$X,
        horizontal = T,
        xlab = nomX,
        main = titre,
        sub = note)
```

Les pays Africains en 2018



Source : Rapport sur le développement humain 2020

- **Commentaire :** La boxplot montre la présence d'au moins quatre valeurs exceptionnelles situées à plus de $1.5 \times (Q3 - Q1)$ de la médiane.

2.2 La distribution de Y

Calculer les paramètres principaux

```
summary(tab$Y)
```

```
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.02423 0.18254 0.39648 1.14048 1.10447 8.09036
```

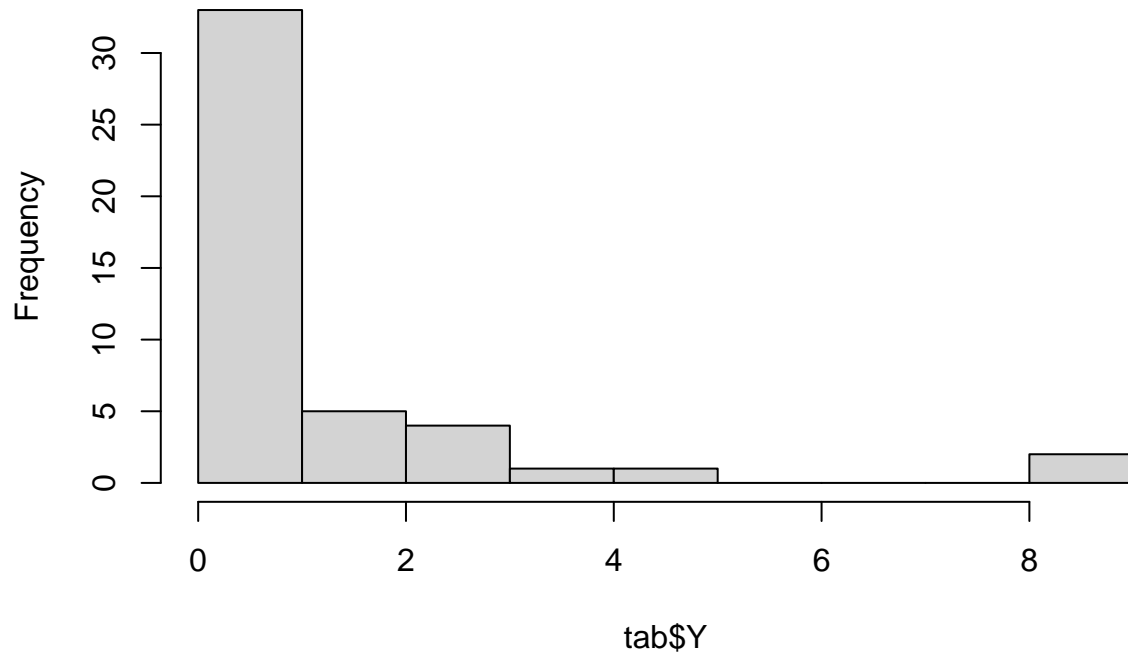
- **Commentaire :** En 2018 les émissions de CO2 des pays d'Afrique varient entre 0.02 t/hab. et 8.1 t/hab. La moyenne est de 1.14 t/hab. La moitié des pays se situent entre 0.18 t/hab (Q1) et 1.10 t/hab (Q3). L'écart entre la moyenne et la médiane suggère une distribution dissymétrique à gauche. Ce que l'on va vérifier avec l'histogramme.

Faire un histogramme

- Histogramme rapide

```
hist(tab$Y)
```

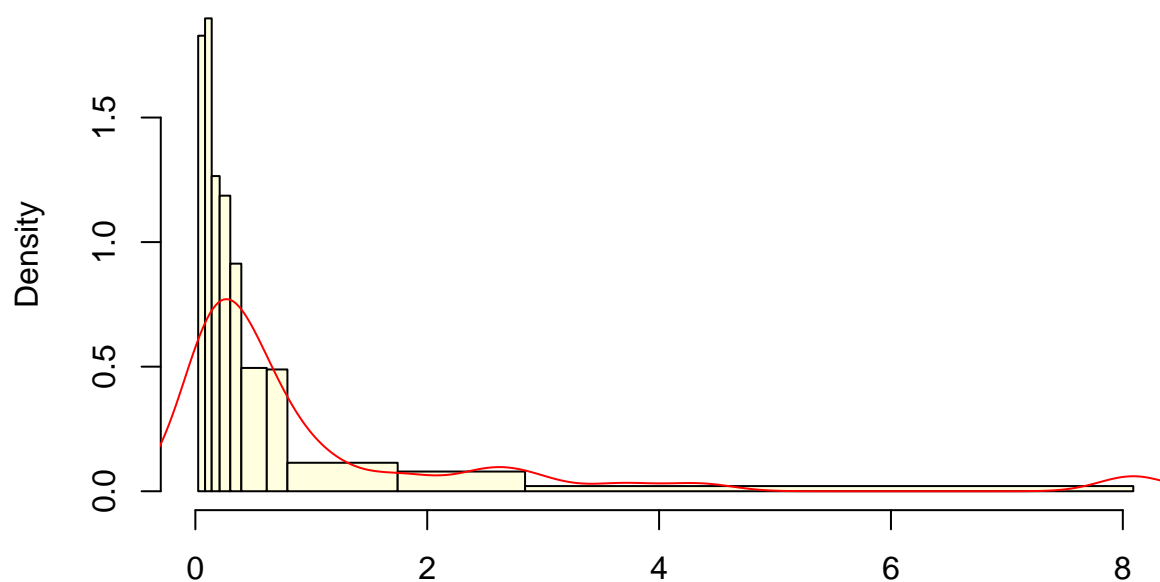
Histogram of tab\$Y



- Histogramme amélioré

```
hist(tab$Y,  
      xlab=nomY,  
      breaks=quantile(tab$Y, c(0,0.1,0.2,0.3,0.4,0.5,0.6,0.7,0.8,0.9,1)),  
      main = titre,  
      sub = note,  
      col="lightyellow")  
lines(density(tab$Y),col="red")
```

Les pays Africains en 2018



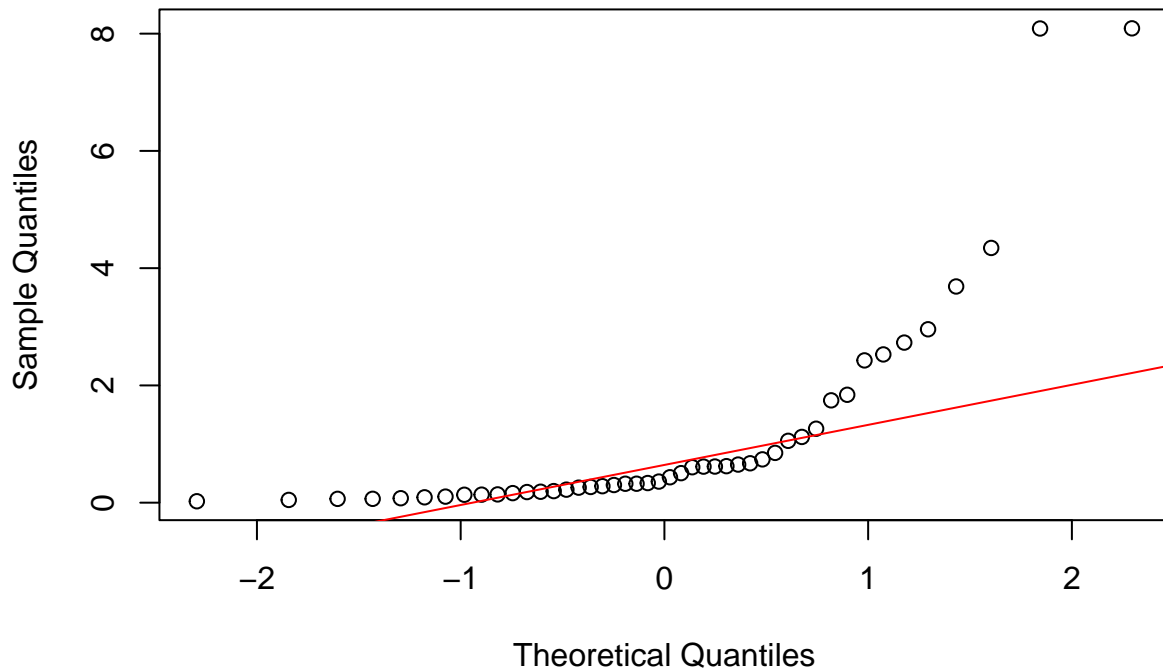
Pollution (t. de CO2/hab).
Source : Rapport sur le développement humain 2020

- **Commentaire :** La distribution de Y est unimodale mais très fortement dissymétrique à gauche.

Tester la normalité

```
# Graphique  
qqnorm(tab$Y)  
qqline(tab$Y, col = "red")
```

Normal Q-Q Plot



```
# test  
shapiro.test(tab$Y)
```

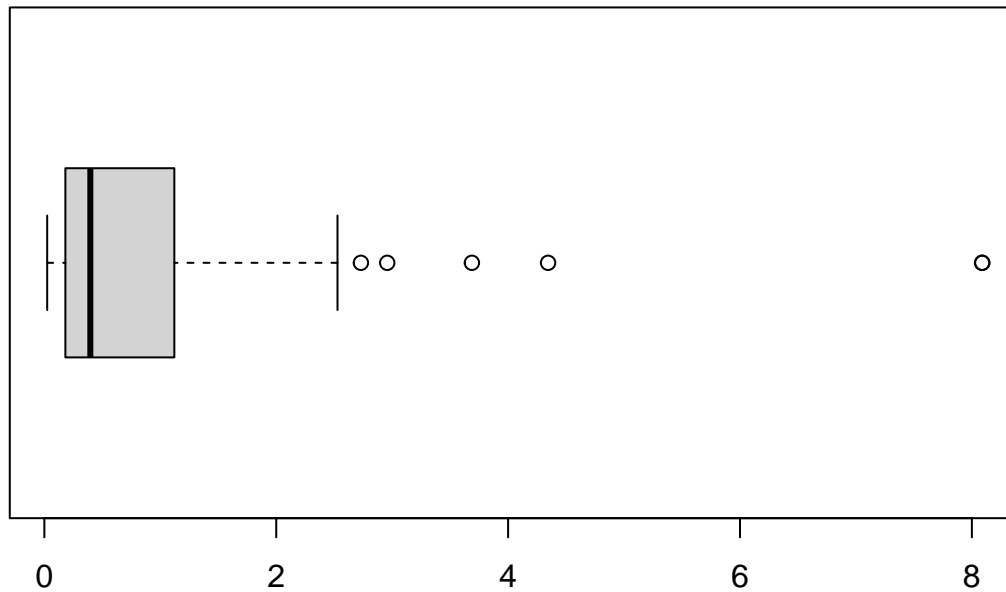
```
##  
##  Shapiro-Wilk normality test  
##  
## data:  tab$Y  
## W = 0.6079, p-value = 7.414e-10
```

- **Commentaire :** Le graphique montre que la distribution ne suit pas une loi gaussienne, ce qui est confirmé par le test de Shapiro-Wilks ($p < 0.001$)

Examiner la présence de valeurs exceptionnelles

```
boxplot(tab$Y,  
        horizontal = T,  
        xlab = nomY,  
        main = titre,  
        sub = note)
```


Les pays Africains en 2018



Pollution (t. de CO2/hab.).

Source : Rapport sur le développement humain 2020

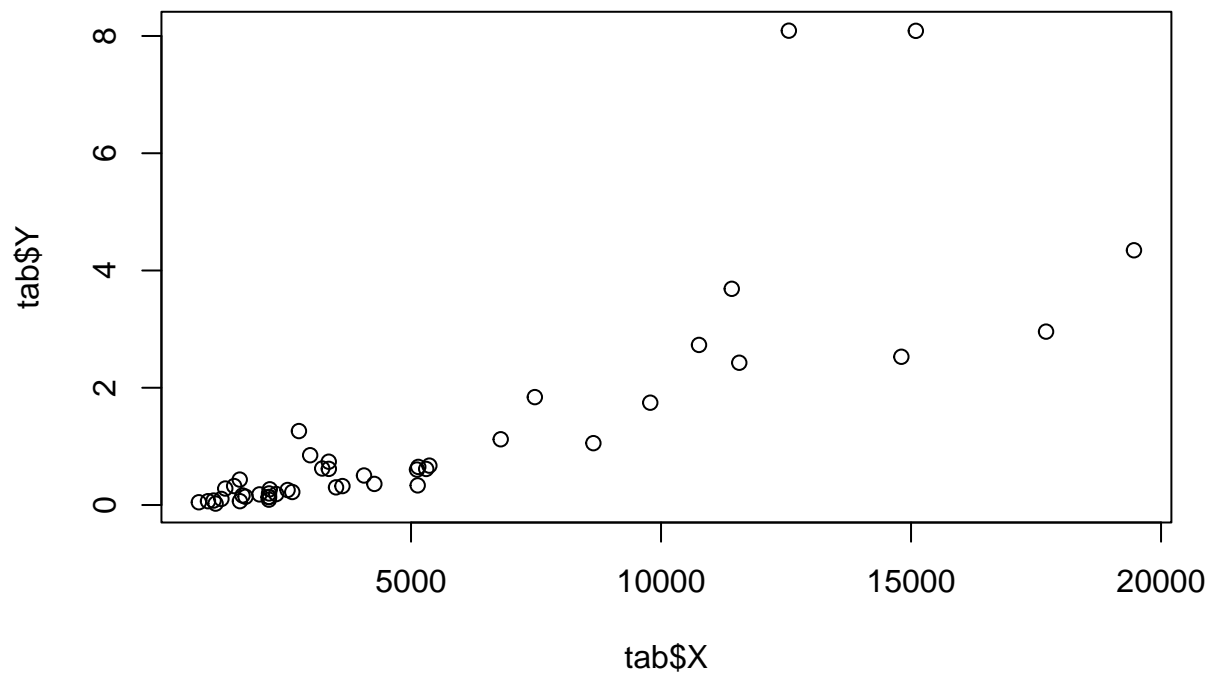
- **Commentaire :** La boxplot montre la présence d'au moins cinq valeurs exceptionnelles situées à plus de $1.5 \times (Q3 - Q1)$ de la médiane.

3. CORRELATION

3.1 Visualiser la relation entre X et Y

- Graphique rapide

```
plot(tab$X,tab$Y)
```

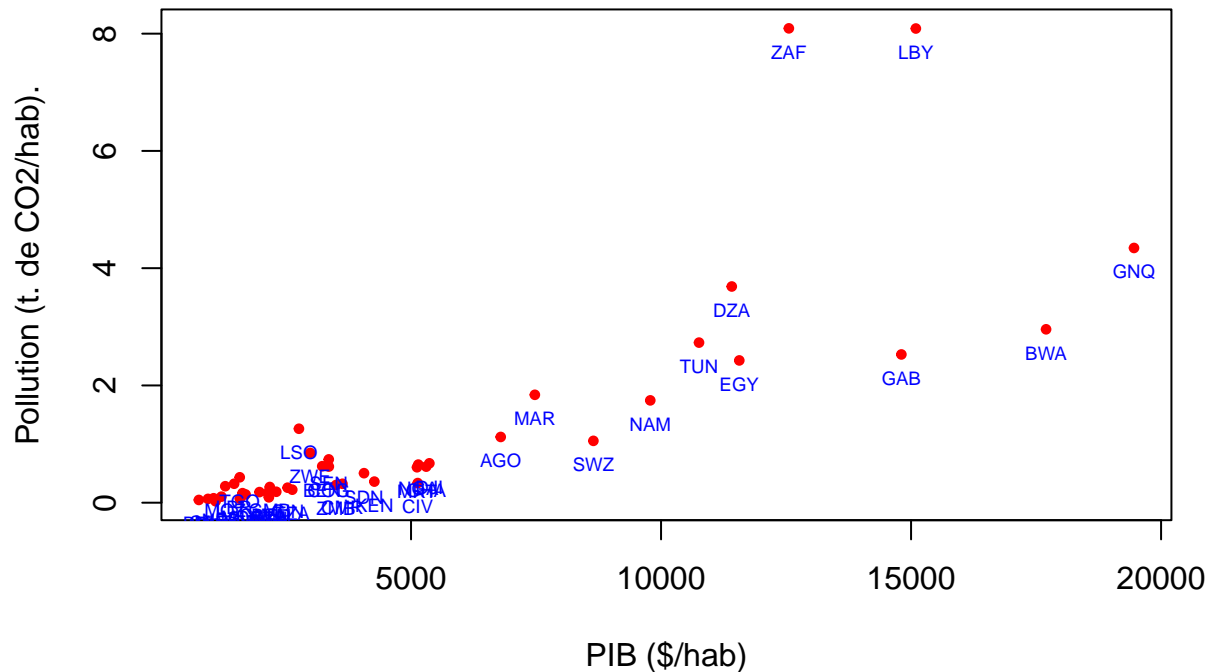


- Graphique amélioré

```
plot(tab$X, tab$Y,
      cex = 0.6,
      pch = 19,
      col = "red",
      xlab = nomX,
      ylab = nomY,
      main = titre,
      sub = note)

text(tab$X, tab$Y, tab$iso3,
      cex = 0.6,
      col = "blue",
      pos = 1)
```

Les pays Africains en 2018



- **Commentaire :** La relation est clairement positive ce qui signifie que plus le PIB/habitant augmente, plus les émissions de CO2 par habitant augmente. Plus un pays est riche, plus il pollue ! Il n'est toutefois pas évident que la relation soit linéaire car deux pays (Afrique du Sud et Libye) s'écartent clairement de la tendance générale et suggèrent une relation de type puissance ou exponentielle.

3.2 Tester la significativité de la relation entre X et Y

Coefficient de Pearson

```
cor.test(tab$X,tab$Y)
```

```
##
## Pearson's product-moment correlation
##
## data: tab$X and tab$Y
## t = 8.9738, df = 44, p-value = 1.688e-11
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
## 0.6703491 0.8873157
## sample estimates:
## cor
## 0.8041573
```

```
cor(tab$X,tab$Y)**2
```

```
## [1] 0.6466689
```

- **Commentaire :** Selon le test du coefficient de Pearson, la relation est très significative ($p < 0.001$) et le pouvoir explicatif de X par rapport à Y (r^2) sera élevé (65%)

Coefficient de Spearman

```
cor.test(tab$X,tab$Y, method = "spearman")

##
## Spearman's rank correlation rho
##
## data: tab$X and tab$Y
## S = 1654, p-value < 2.2e-16
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##      rho
## 0.8979957
```

- **Commentaire :** Le coefficient de corrélation de Spearman (+0.90) est sensiblement plus élevée que celui de Pearson (+0.80). Ceci constitue un signal d'alerte et suggère (i) soit la présence de valeurs exceptionnelles, (ii) soit l'existence d'une relation non linéaire.

4. REGRESSION LINEAIRE

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

```
monmodel <- lm(tab$Y~tab$X)
summary(monmodel)

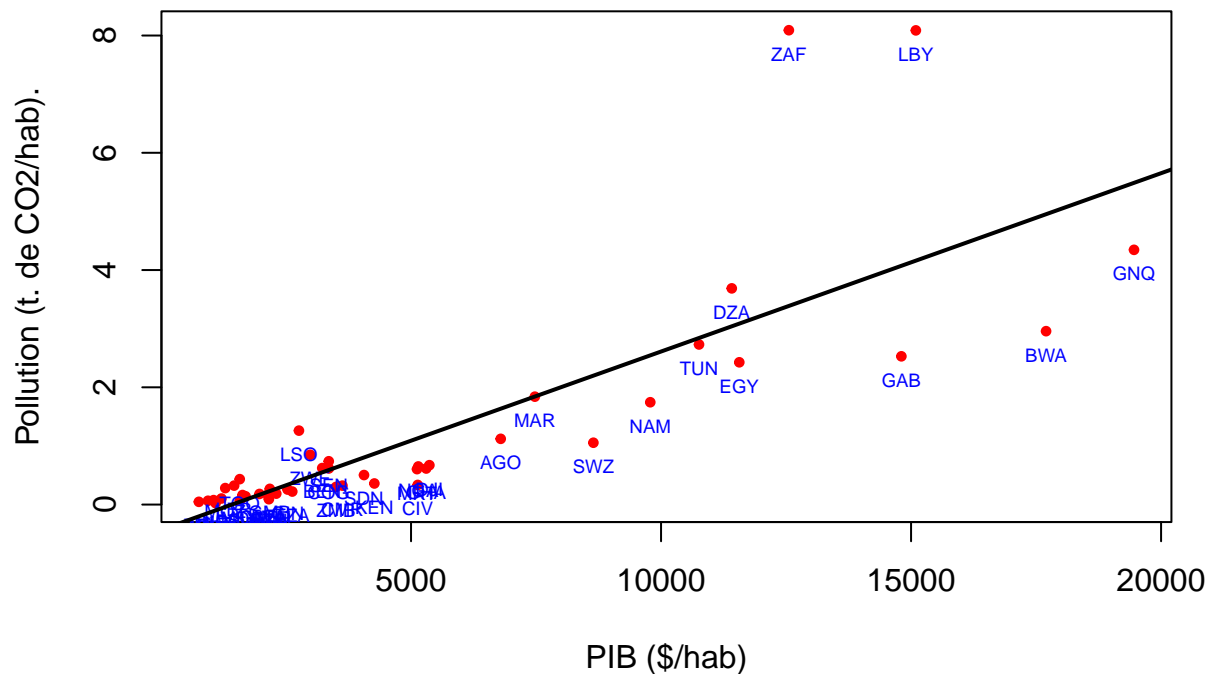
##
## Call:
## lm(formula = tab$Y ~ tab$X)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.9949 -0.5007 -0.0551  0.1633  4.7028
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.4317539   0.2375862  -1.817   0.076 .
## tab$X        0.0003042   0.0000339   8.974 1.69e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.088 on 44 degrees of freedom
## Multiple R-squared:  0.6467, Adjusted R-squared:  0.6386
## F-statistic: 80.53 on 1 and 44 DF,  p-value: 1.688e-11
```

- **Commentaire :** L'équation de la droite est donc $*Y = 0.0003*X - 0.432*$. Le coefficient de pente de la droite indique que les émissions de CO2 augmentent de 0.0003 tonnes chaque fois que le PIB par habitant augmente de 1 dollar. Ou si l'on préfère, que les émissions de CO2 augmentent de 0.3 tonnes chaque fois que le PIB/hab. augmente de 1000 dollars. La constante (Intercept) indique la valeur qui correspondrait à un pays totalement pauvre et elle serait négative ce qui est évidemment absurde. Le modèle linéaire peut aboutir à des absurdités ...

4.2 Visualiser la droite

```
plot(tab$X,tab$Y,
     cex = 0.6,
     pch = 19,
     col = "red",
     xlab = nomX,
     ylab = nomY,
     main = titre,
     sub = note)
text(tab$X, tab$Y, tab$iso3,
     cex = 0.6,
     col = "blue",
     pos = 1)
abline(monmodel, col ="black", lwd =2)
```

Les pays Africains en 2018



Source : Rapport sur le développement humain 2020

Commentaire: La droite s'ajuste plus ou moins au nuage de points mais on remarque que les résidus sont mal répartis autour de celle-ci (*autocorrélation*) et que les points s'éloignent de plus en plus de la droite au fur et à mesure que X augmente ce qui signifie que la variance n'est pas constante (*hétéroscédasticité*). Même s'il semble avoir un fort pouvoir explicatif, le modèle semble donc souffrir de défauts importants que l'on discutera dans la partie finale.

4.3 Calculer les valeurs estimées et les résidus

```
# Extraction des valeurs estimées et résiduelles
tab$Yest <- monmodel$fitted.values
tab$Yres <- monmodel$residuals

# Affichage du tableau trié
tab[order(tab$Yres),]
```

##	iso3	name	X	Y	Yest	Yres
## 5	BWA	Botswana	17700.3152	2.95736420	4.95231246	-1.99494826
## 16	GAB	Gabon	14806.5905	2.52836465	4.07210181	-1.54373716
## 41	SWZ	Swaziland	8647.0887	1.05493272	2.19850997	-1.14357725
## 21	GNQ	Eq. Guinea	19458.9245	4.34492637	5.48724470	-1.14231833
## 32	NAM	Namibia	9784.5769	1.74504687	2.54451013	-0.79946325
## 7	CIV	Côte d'Ivoire	5133.5905	0.33469716	1.12977713	-0.79507998
## 13	EGY	Egypt	11564.7950	2.42640693	3.08601530	-0.65960837
## 17	GHA	Ghana	5303.5336	0.61471889	1.18147028	-0.56675138
## 11	DJI	Djibouti	5366.7107	0.67179534	1.20068744	-0.52889210
## 30	MRT	Mauritania	5119.7260	0.60443125	1.12555986	-0.52112861
## 1	AGO	Angola	6793.7085	1.12098146	1.63475038	-0.51376892
## 22	KEN	Kenya	4266.8368	0.35970573	0.86612877	-0.50642305
## 34	NGA	Nigeria	5145.2871	0.64987572	1.13333499	-0.48345928
## 8	CMR	Cameroon	3628.1177	0.32269646	0.67184374	-0.34914729
## 48	ZMB	Zambia	3500.5119	0.30129599	0.63302873	-0.33173275
## 36	SDN	Sudan	4059.5331	0.50347675	0.80307131	-0.29959456
## 45	TZA	Tanzania	2625.1980	0.22195235	0.36677650	-0.14482415
## 35	RWA	Rwanda	2157.3935	0.09123742	0.22448015	-0.13324273
## 44	TUN	Tunisia	10759.6827	2.73031139	2.84111696	-0.11080557
## 15	ETH	Ethiopia	2161.6109	0.13674141	0.22576299	-0.08902158
## 46	UGA	Uganda	2151.6788	0.13506038	0.22274186	-0.08768148
## 28	MLI	Mali	2305.2361	0.18661531	0.26945079	-0.08283548
## 18	GIN	Guinea	2531.2341	0.25609996	0.33819464	-0.08209468
## 4	BFA	Burkina Faso	2160.5895	0.19739801	0.22545229	-0.02805428
## 26	MAR	Morocco	7476.1792	1.84045155	1.84234373	-0.00189218
## 20	GNB	Guinea-Bissau	1969.2724	0.18117668	0.16725762	0.01391906
## 42	TCO	Chad	1577.9727	0.06553287	0.04823243	0.01730044
## 10	COG	Congo	3356.2418	0.61610961	0.58914478	0.02696483
## 19	GMB	Gambia	2175.5690	0.26780913	0.23000875	0.03780039
## 39	SLE	Sierra Leone	1690.8316	0.14093841	0.08256175	0.05837667
## 3	BEN	Benin	3224.0433	0.62247911	0.54893275	0.07354636
## 27	MDG	Madagascar	1629.6623	0.16302068	0.06395532	0.09906535
## 9	COD	Dem. Rep. Congo	1091.9213	0.02422917	-0.09961426	0.12384343
## 38	SEN	Senegal	3354.8272	0.73865658	0.58871450	0.14994208
## 33	NER	Niger	1207.7762	0.10333371	-0.06437363	0.16770734
## 31	MWI	Malawi	1051.1249	0.07611109	-0.11202366	0.18813475
## 6	CAF	Central African Rep.	938.9888	0.06508274	-0.14613312	0.21121586
## 2	BDI	Burundi	756.5941	0.04667763	-0.20161380	0.24829143
## 23	LBR	Liberia	1462.4118	0.32340916	0.01308122	0.31032794
## 29	MOZ	Mozambique	1284.9965	0.28091941	-0.04088480	0.32180421
## 49	ZWE	Zimbabwe	2982.9890	0.84928791	0.47560907	0.37367884
## 43	TGO	Togo	1574.2385	0.43325640	0.04709655	0.38615984
## 12	DZA	Algeria	11414.5995	3.68769052	3.04032896	0.64736157
## 25	LSO	Lesotho	2758.1290	1.26133125	0.40721135	0.85411991
## 24	LYB	Libya	15096.0769	8.08792735	4.16015754	3.92776981
## 47	ZAF	South Africa	12556.2802	8.09035696	3.38760439	4.70275256

- **Commentaire :** Le tableau permet de repérer les pays qui s'éloignent le plus de la droite en raison d'une surestimation ou d'une sous-estimation de leurs émissions de CO2 par le PIB. Les résidus négatifs correspondent à des pays qui émettent moins de CO2 que ce que laisserait prévoir leur PIB. C'est par exemple le cas du Botswana dont le PIB élevé (17700\$/hab.) laissait prévoir 4.95 t. de CO2 par habitant mais qui en pratique n'en émet que 2.96 soit un résidu de -2 tonnes. Inversement le PIB de l'Afrique du Sud (12256 \$/hab) laissait prévoir 3.9 tonnes de CO2 par habitant alors que la valeur observée est de

8.1 tonnes, soit un résidu de +4.7 tonnes de plus que prévu. Dans les deux cas on peut chercher des explications ad hoc (e.g. importance de la production de charbon en Afrique du Sud) mais il faut aussi se demander si ces écarts ne sont pas justes liés à une mauvaise spécification de notre modèle ...

4.4 Sauvegarder les résultats du modèle

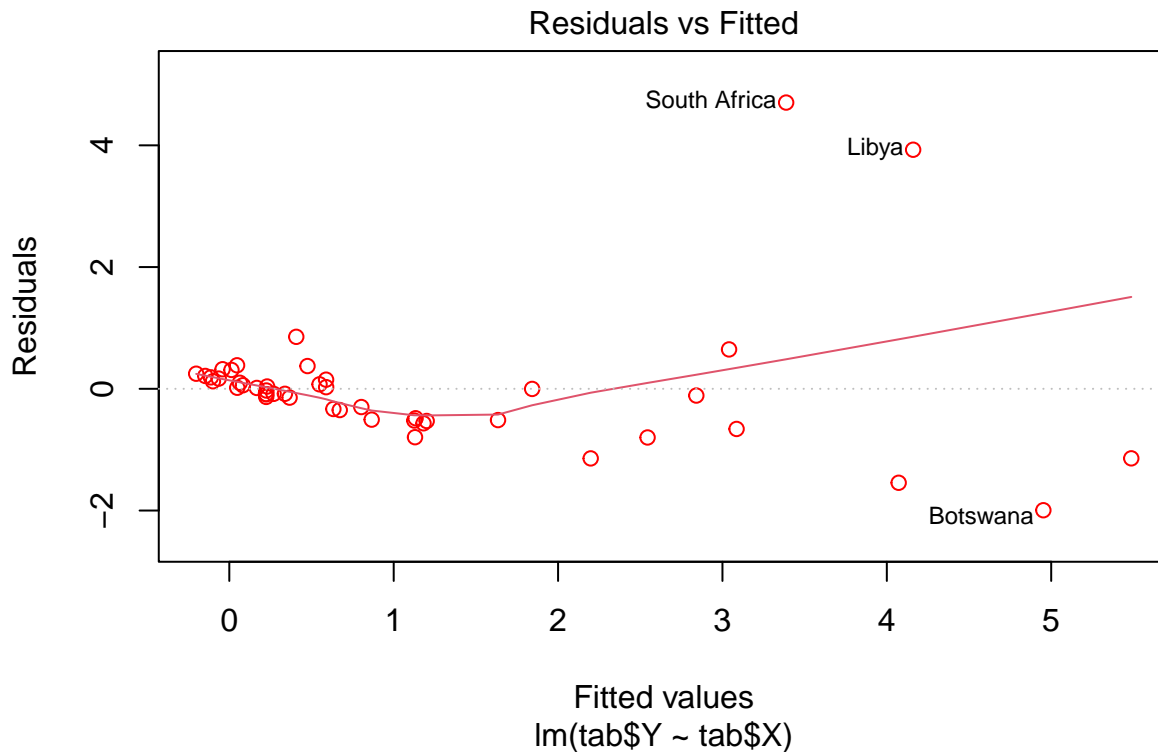
```
write.table(x = tab,  
           file = "result.csv",  
           row.names = FALSE)
```

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

```
plot(monmodel,  
     which = 1,  
     labels.id = tab$name,  
     col="red")
```



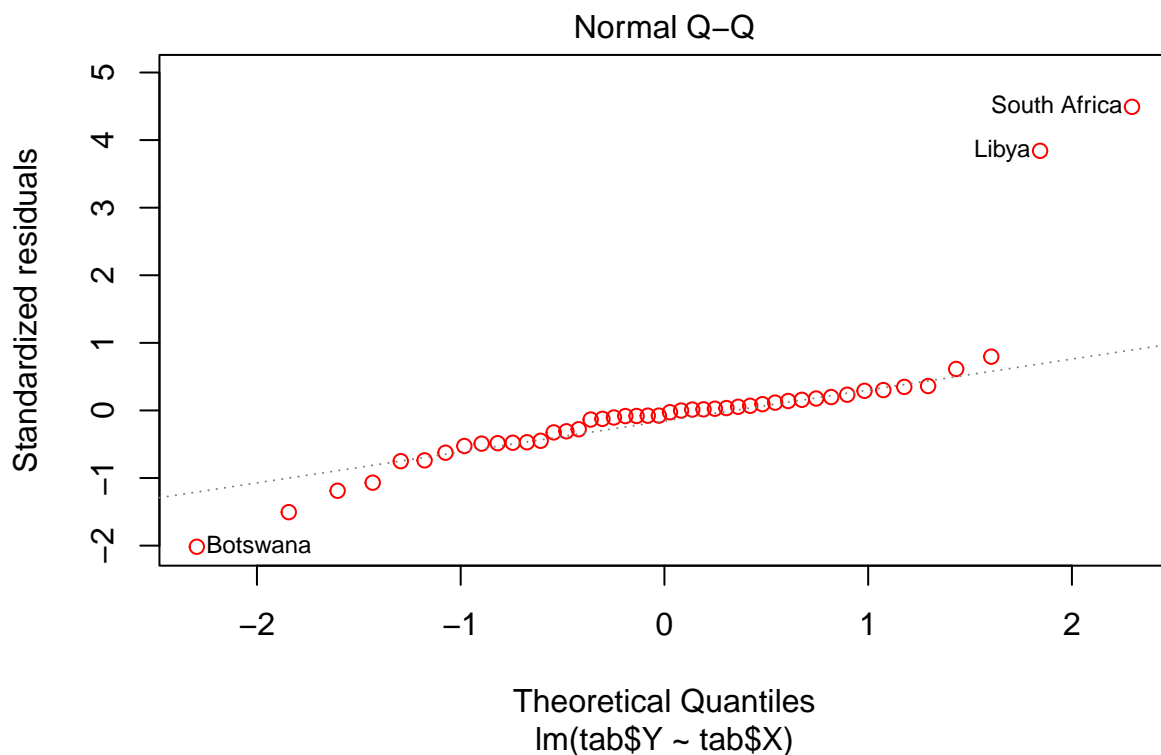
```
durbinWatsonTest(monmodel)
```

```
## lag Autocorrelation D-W Statistic p-value  
## 1 0.04054838 1.91116 0.714  
## Alternative hypothesis: rho != 0
```

- **Commentaire** : le graphique permet de voir que les résidus ne sont pas indépendants des valeurs estimées de Y, ce qui signifie que les points se situent en moyenne tantôt au dessus de la droite de régression, tantôt en dessous ce qui fausse leur estimation. Dans un modèle sans autocorrélation, la courbe rouge devrait suivre la ligne pointillée correspondant à une moyenne nulle des résidus, ce qui n'est visiblement pas le cas. On peut s'en assurer à l'aide du *test de Durbin Watson* qui pose l'hypothèse H_0 : *Il existe une autocorrélation des résidus*. Cette hypothèse ne peut pas être rejetée ($p > 0.66$) donc il existe bien une autocorrélation des résidus qui va fausser les prévisions du modèle de régression linéaire.

5.2 Normalité des résidus

```
plot(monmodel,
     which = 2,
     labels.id = tab$name,
     col="red")
```



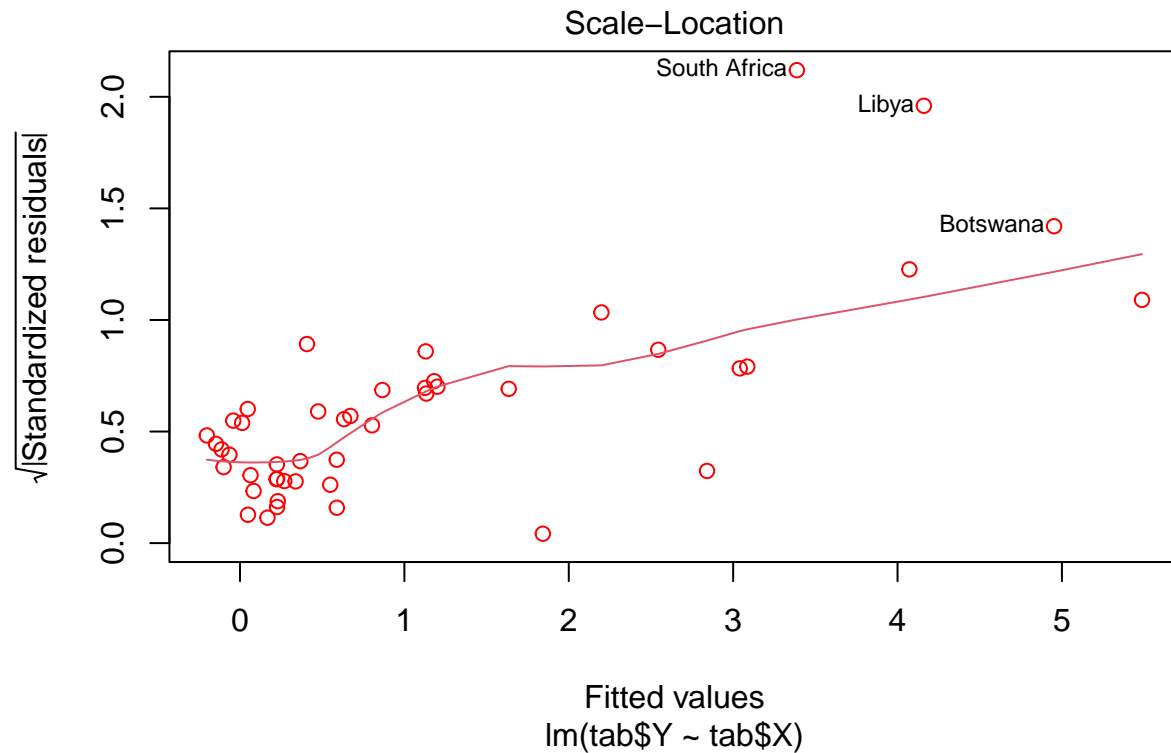
```
shapiro.test(tab$Yres)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  tab$Yres
## W = 0.68437, p-value = 1.163e-08
```

- **Commentaire** : La normalité de la distribution des résidus est également une condition importante de validité du modèle de régression linéaire puisqu'elle permet de définir un intervalle de confiance des estimations en se servant de l'écart-type de ces résidus (e.g. + ou - 2 écarts-type pour un intervalle de confiance à 95%). Mais il est clair ici au vu du diagramme QQ plot que la condition de normalité des résidus n'est pas vérifiée, ce que confirme le test de shapiro ($p < 0.001$)

5.3 Homogénéité des résidus

```
plot(monmodel,
     which = 3,
     labels.id = tab$name,
     col="red")
```



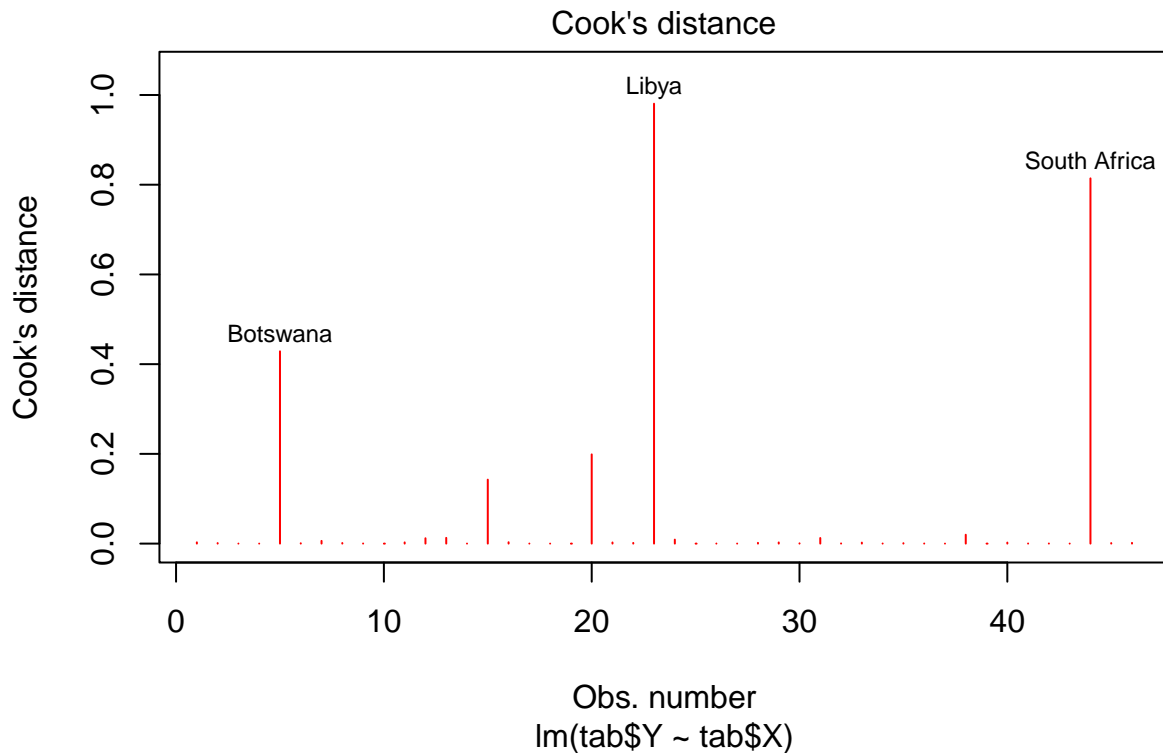
```
ncvTest(monmodel)
```

```
## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 65.47898, Df = 1, p = 5.8737e-16
```

- **Commentaire** : En liaison avec ce qui précède, l'analyse de l'homogénéité des résidus permet de vérifier si la variance des résidus est constante et donc si l'intervalle de confiance sera le même pour l'ensemble des valeurs estimées. Ici, ce n'est clairement pas le cas puisque le graphique montre un net accroissement de la variance des résidus lorsque la valeur à estimer augmente. On peut vérifier l'absence d'homogénéité (appelée *hétéroscédasticité*) en appliquant le *test de Breush-Pagan* qui examine l'hypothèse "H0 : la distribution des résidus est homogène". Dans notre exemple H0 est rejetée ($p < 0.001$) ce qui signifie que l'hypothèse d'homogénéité est clairement violée.

5.4 Absence de valeurs exceptionnellement influentes

```
plot(monmodel,
     which = 4,
     labels.id = tab$name,
     col="red")
```



```
outlierTest(monmodel, labels = tab$name)
```

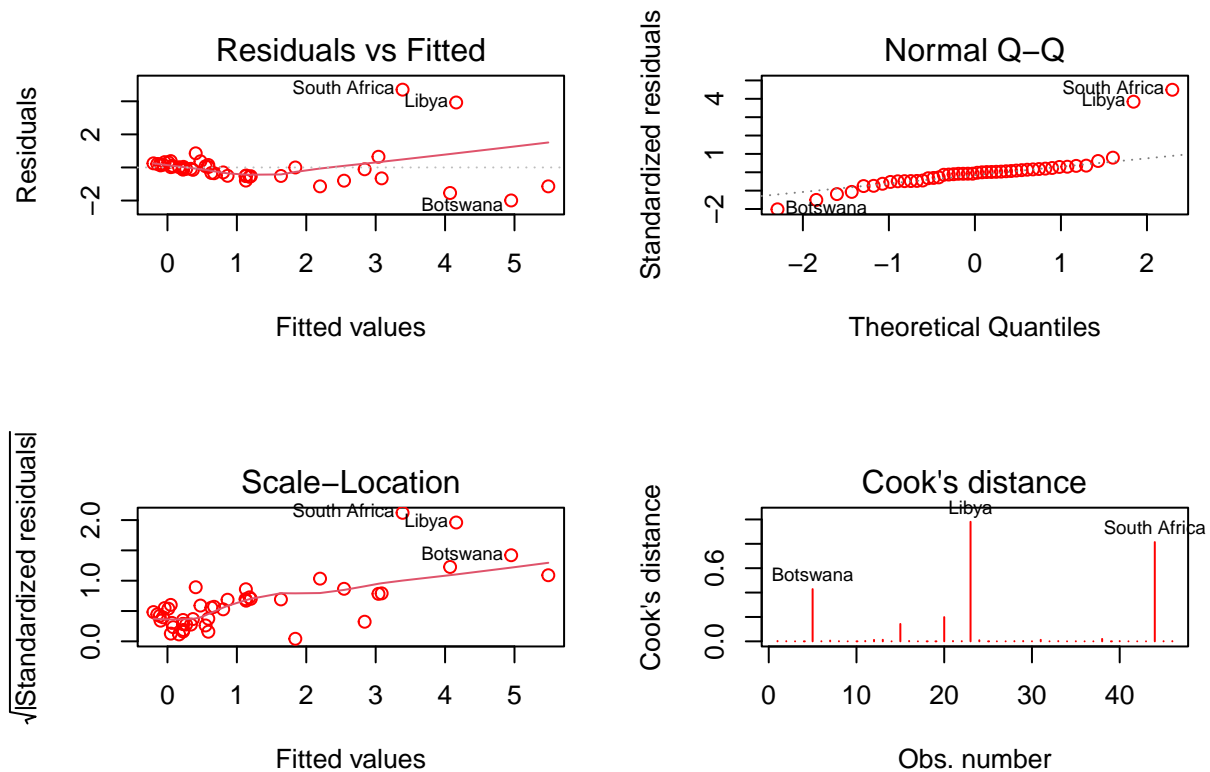
```
##          rstudent unadjusted p-value Bonferroni p
## South Africa 6.034761      3.2546e-07  1.4971e-05
## Libya       4.657740      3.0790e-05  1.4164e-03
```

- **Commentaire :** Le dernier test consiste à vérifier si la relation observée est bien le résultat d'un ensemble d'observations indépendante et non pas l'effet de la présence d'une ou deux valeurs exceptionnelles. Plusieurs tests sont ici possibles qui visent au même objectif : déterminer à quel point le retrait d'une valeur unique modifie le résultat de l'analyse, c'est à dire le coefficient de détermination et les paramètres a et b de l'équation $Y=aX+b$. Le graphique proposé par R utilise la *distance de Cook* pour mettre en valeur l'influence potentielle des valeurs exceptionnelles et on y retrouve sans surprise la Libye, l'Afrique du Sud et le Botswana. On peut arriver à un résultat similaire en utilisant le test de Bonferroni qui signale le caractère exceptionnellement influent de l'Afrique du Sud et de la Libye.

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 :

```
par(mfrow=c(2,2))
plot(monmodel,
     which = c(1,2,3,4),
     labels.id = tab$name,
     col="red")
```



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

```
tab2<-tab[!(tab$iso3 %in% c("ZAF","BWA","LBY")),]
```

Corrélation

```
cor.test(tab2$X,tab2$Y, method="pearson")

##
## Pearson's product-moment correlation
##
## data: tab2$X and tab2$Y
## t = 16.456, df = 41, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
## 0.8770905 0.9627927
## sample estimates:
```

```
##          cor
## 0.9319357
cor.test(tab2$X,tab2$Y, method="spearman")

##
## Spearman's rank correlation rho
##
## data:  tab2$X and tab2$Y
## S = 1614, p-value < 2.2e-16
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##          rho
## 0.8781335
```

Régression

```
monmodel2 <- lm(tab2$Y~tab2$X)
summary(monmodel2)

##
## Call:
## lm(formula = tab2$Y ~ tab2$X)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.67137 -0.21168 -0.02265  0.12596  1.33042
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.452e-01  8.348e-02  -2.937  0.00542 **
## tab2$X       2.280e-04  1.385e-05  16.456 < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3666 on 41 degrees of freedom
## Multiple R-squared:  0.8685, Adjusted R-squared:  0.8653
## F-statistic: 270.8 on 1 and 41 DF,  p-value: < 2.2e-16

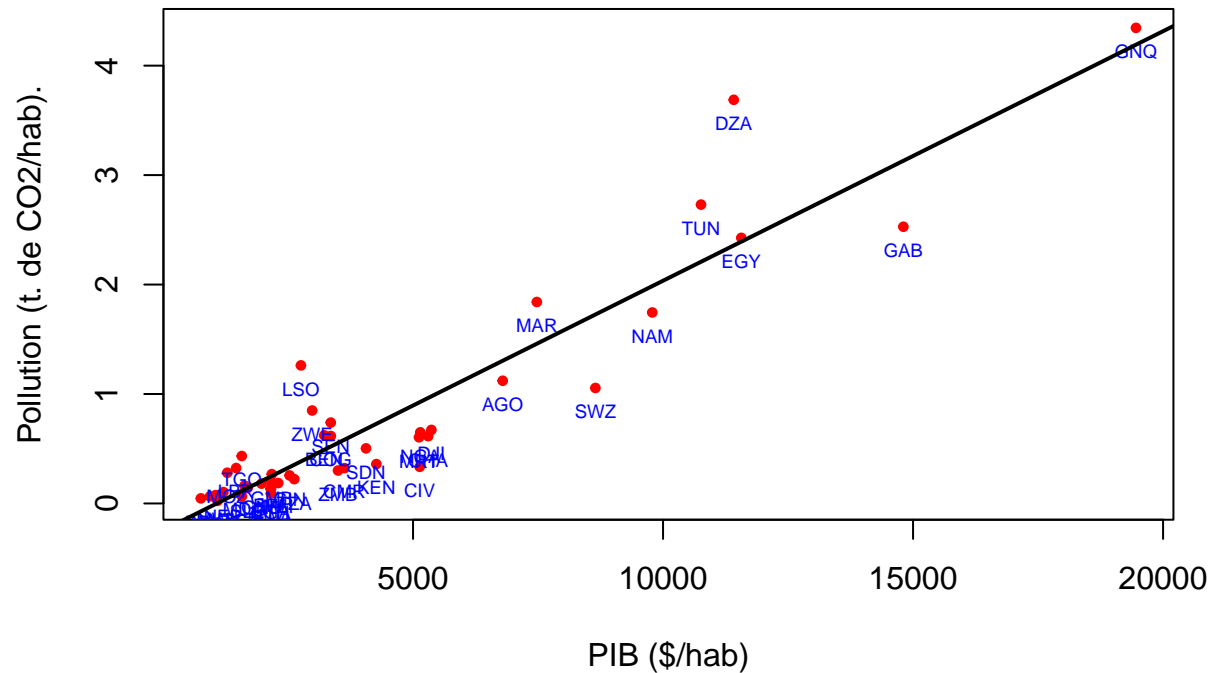
# Extraction des valeurs estimées et résiduelles
tab2$Yest <- monmodel2$fitted.values
tab2$Yres <- monmodel2$residuals
```

Visualisation

```
plot(tab2$X,tab2$Y,
     cex = 0.6,
     pch = 19,
     col = "red",
     xlab = nomX,
     ylab = nomY,
     main = titre,
     sub = note)
text(tab2$X, tab2$Y, tab2$iso3,
     cex = 0.6,
```

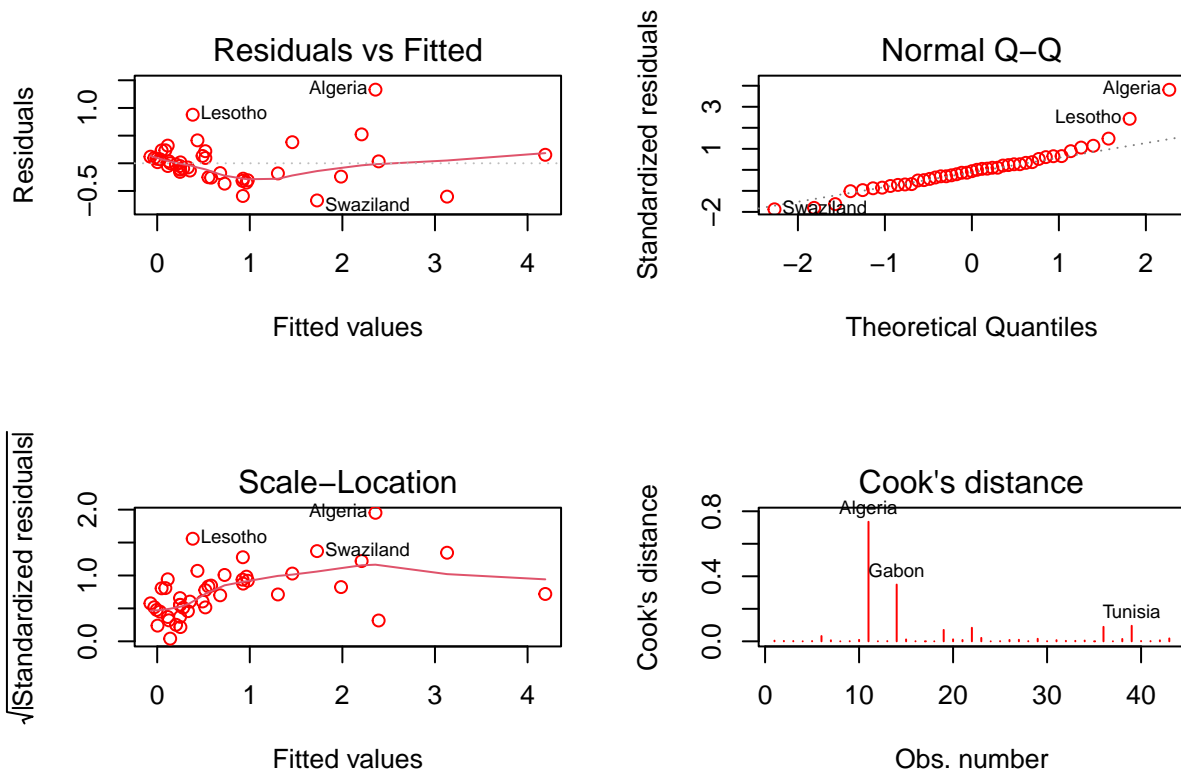
```
col = "blue",
pos = 1)
abline(monmodel2, col = "black", lwd = 2)
```

Les pays Africains en 2018



Diagnostics

```
par(mfrow=c(2,2))
plot(monmodel2,
     which = c(1,2,3,4),
     labels.id = tab2$name,
     col="red")
```



```
durbinWatsonTest(monmodel12)
```

```
## lag Autocorrelation D-W Statistic p-value
## 1 0.05326152 1.856255 0.602
## Alternative hypothesis: rho != 0
```

```
shapiro.test(tab2$Yres)
```

```
##
## Shapiro-Wilk normality test
##
## data: tab2$Yres
## W = 0.91406, p-value = 0.003438
```

```
ncvTest(monmodel12)
```

```
## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 14.98559, Df = 1, p = 0.00010834
```

```
outlierTest(monmodel12, labels = tab2$name)
```

```
##          rstudent unadjusted p-value Bonferroni p
## Algeria 4.679998      3.263e-05    0.0014031
```

- **Commentaire :** Le nouveau modèle affiche une corrélation beaucoup plus élevée ($r = 0.96$) et une bien meilleure qualité d'ajustement ($r^2 = 86.5\%$). Il demeure une forte autocorrélation des résidus ($p > 0.60$) mais les résidus sont à peu près gaussiens ($p > 0.05$). L'hétéroscédasticité demeure élevée ($p < 0.001$) et on trouve une nouvelle valeur exceptionnellement influente (Algérie). Il y a donc d'indéniables progrès mais le modèle n'est pas encore tout à fait satisfaisant.

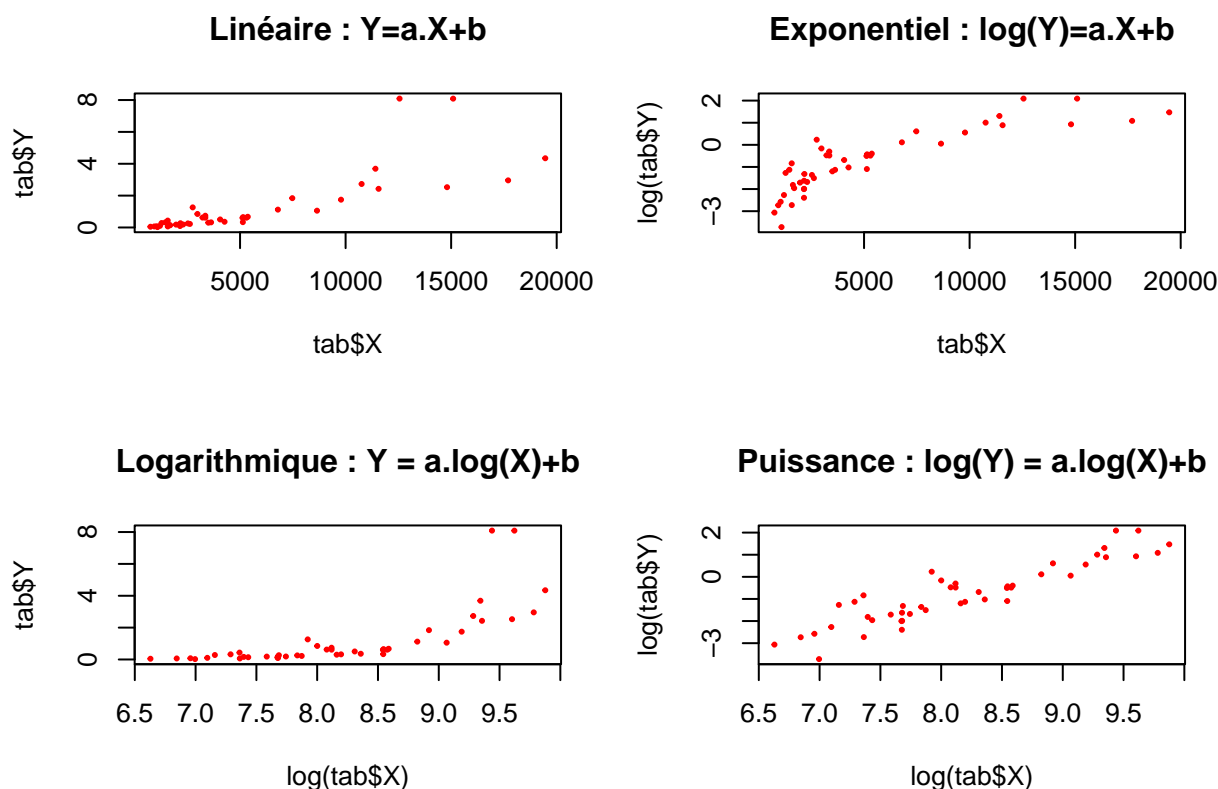
6.2 Modèles non linéaires

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

```
par(mfrow=c(2,2))

plot(tab$X,tab$Y, main = "Linéaire : Y=a.X+b", pch=20, col="red",cex=0.5)
plot(tab$X,log(tab$Y), main = "Exponentiel : log(Y)=a.X+b", pch=20, col="red",cex=0.5)
plot(log(tab$X),tab$Y, main = "Logarithmique : Y = a.log(X)+b", pch=20, col="red",cex=0.5)
plot(log(tab$X),log(tab$Y), main = "Puissance : log(Y) = a.log(X)+b", pch=20, col="red",cex=0.5)
```



- **Commentaire :** Un simple examen visuel laisse présager que le modèle puissance est celui qui s'ajustera le mieux à une droite et offrira une répartition régulière des résidus conforme aux hypothèses.

Calcul des coefficients de corrélation

```
paste("Linéaire : ",round(cor(tab$X,tab$Y),3))

## [1] "Linéaire : 0.804"

paste("Exponentiel : ", round(cor(tab$X,log(tab$Y)),3))

## [1] "Exponentiel : 0.85"
```

```
paste("Logarithmique : ",round(cor(log(tab$X),tab$Y),3))
```

```
## [1] "Logarithmique : 0.722"
```

```
paste("Puissance : ", round(cor(log(tab$X),log(tab$Y)),3))
```

```
## [1] "Puissance : 0.913"
```

- **Commentaire :** Le calcul des coefficients de corrélation confirme que cette solution donne le meilleur ajustement aux données. Noter bien que ce critère ne suffit pas à lui seul à choisir un modèle. Un modèle qui aurait un meilleur ajustement mais violerait les hypothèses ne devrait pas être retenu face à un modèle ayant un ajustement plus faible mais des résidus mieux distribués.

Préparation des données

On crée un nouveau tableau de données

```
don$X<-log(don$PIB)
don$Y<-log(don$CO2)
tab3<-don[,c("iso3","name","X","Y")]
tab3<-tab3[complete.cases(tab3), ]
nomXlog <- "log(PIB en $/hab)"
nomYlog <- "log(CO2 en t./hab)"
titre <- "Les pays Africains en 2018"
note <- "Source : Rapport sur le développement humain 2020"
```

Régression

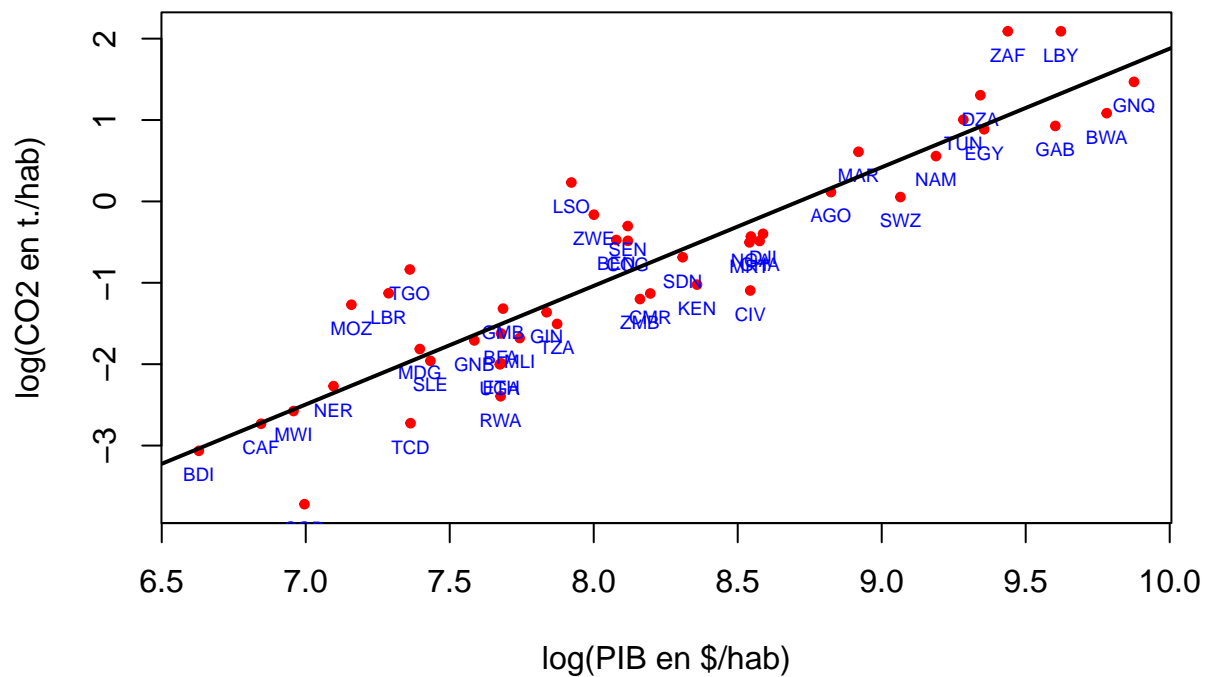
```
monmodel3 <- lm(tab3$Y~tab3$X)
summary(monmodel3)
```

```
##
## Call:
## lm(formula = tab3$Y ~ tab3$X)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.21875 -0.34992 -0.09064  0.27574  1.38323
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -12.69651    0.80784  -15.72  <2e-16 ***
## tab3$X       1.45733    0.09822   14.84  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5672 on 44 degrees of freedom
## Multiple R-squared:  0.8334, Adjusted R-squared:  0.8296
## F-statistic: 220.2 on 1 and 44 DF, p-value: < 2.2e-16
# Extraction des valeurs estimées et résiduelles
tab3$Yest <- monmodel$fitted.values
tab3$Yres <- monmodel$residuals
```


Visualisation

```
plot(tab3$X,tab3$Y,
     cex = 0.6,
     pch = 19,
     col = "red",
     xlab = nomXlog,
     ylab = nomYlog,
     main=titre,
     sub = note)
text(tab3$X, tab3$Y, tab3$iso3,
     cex = 0.6,
     col = "blue",
     pos = 1)
abline(monmodel3, col ="black", lwd =2)
```

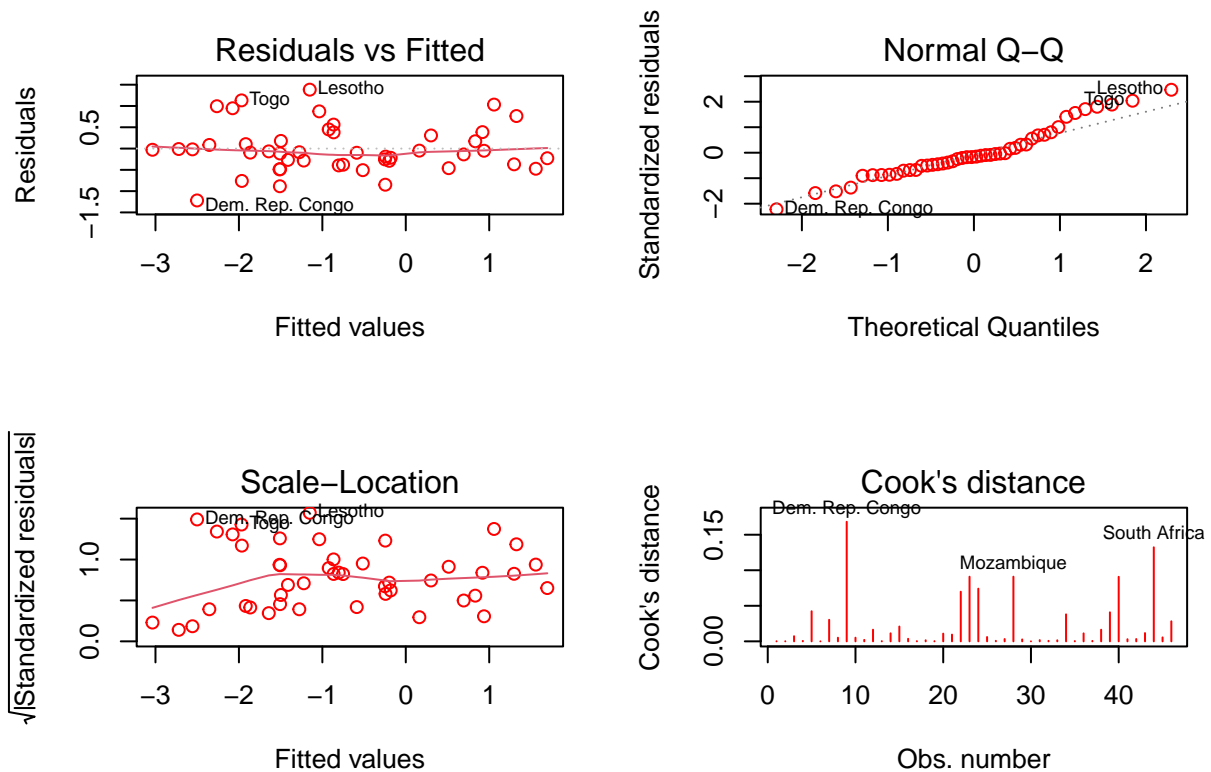
Les pays Africains en 2018



Source : Rapport sur le développement humain 2020

Diagnostics

```
par(mfrow=c(2,2))
plot(monmodel3,
     which = c(1,2,3,4),
     labels.id = tab3$name,
     col="red")
```



```
durbinWatsonTest(monmodel3)
```

```
## lag Autocorrelation D-W Statistic p-value
## 1 0.03349644 1.87894 0.7
## Alternative hypothesis: rho != 0
```

```
shapiro.test(tab3$Yres)
```

```
##
## Shapiro-Wilk normality test
##
## data: tab3$Yres
## W = 0.68437, p-value = 1.163e-08
```

```
ncvTest(monmodel3)
```

```
## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 1.248169, Df = 1, p = 0.2639
```

```
outlierTest(monmodel3, labels = tab3$name)
```

```
## No Studentized residuals with Bonferroni p < 0.05
## Largest |rstudent|:
##      rstudent unadjusted p-value Bonferroni p
## Lesotho 2.628744 0.011838 0.54453
```

```
exp(-12.69)
```

```
## [1] 3.08179e-06
```

- **Commentaires :** Outre sa qualité d'ajustement élevée ($r^2 = 83\%$), le modèle final respecte beaucoup mieux les hypothèses théoriques d'un modèle de régression linéaire. Il demeure certes une légère

autocorrélation des résidus et une distribution qui n'est pas tout à fait gaussienne. Mais les résidus sont désormais homogènes ($p > 0.26$) et aucune valeur influente n'est plus détectée par le test de Bonferoni. Bref, le modèle est acceptable.

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

Le modèle ayant été ajusté sous forme bi-logarithmique, il faut en rétablir l'équation sous la forme $Y = f(X)$, ce qui suppose de transformer l'équation de la façon suivante :

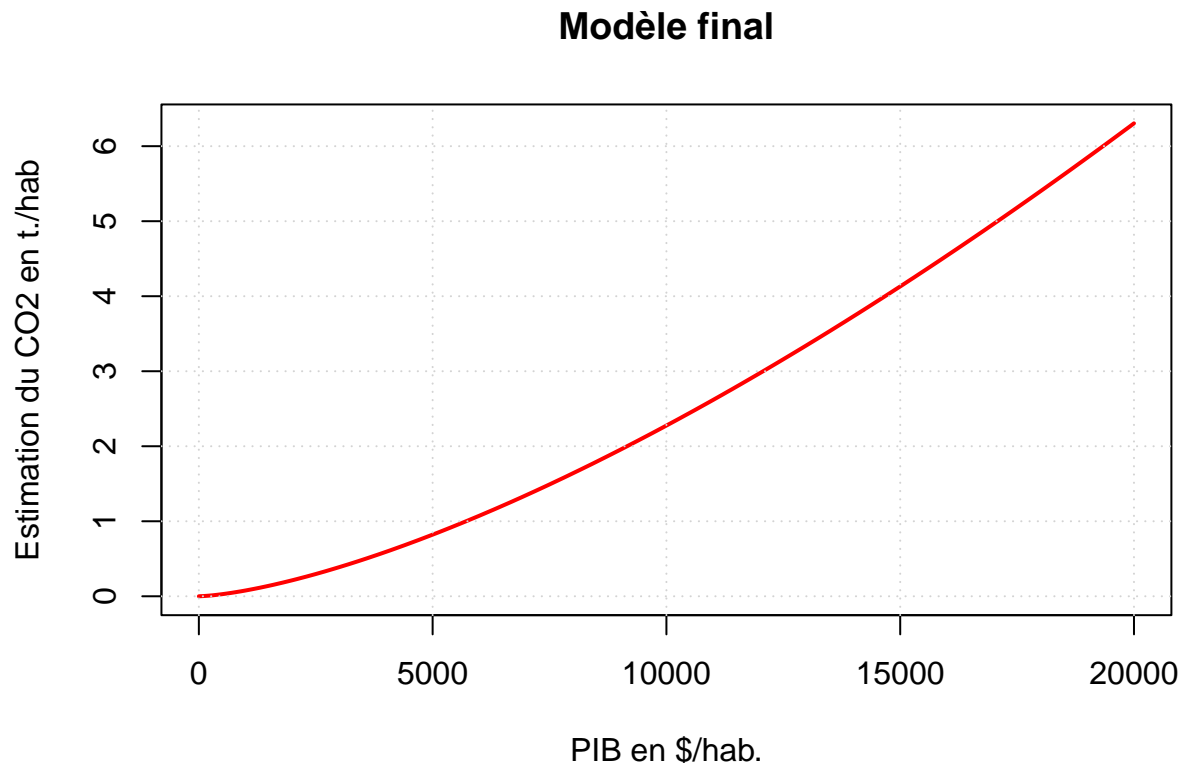
- $\log(Y) = a \times \log(X) + b \Leftrightarrow Y = e^b \times X^a$

Ce qui nous donne l'équation finale :

- $\log(CO2) = a \times \log(PIB) + b \Leftrightarrow CO2 = e^{-12.696} \times PIB^{1.47} \Leftrightarrow CO2 = 0.000003 \times PIB^{1.47}$

Que l'on peut représenter de la façon suivante :

```
x<-seq(0,20000,100)
y<- 0.000003*(x**1.47)
plot(x,y,
      type="l",
      col="red",
      lwd =2,
      xlab = "PIB en $/hab.",
      ylab = "Estimation du CO2 en t./hab",
      main = "Modèle final")
grid()
```



- Commentaire :** Notre modèle final offre une représentation assez fiable de la relation qui existe entre le PIB par habitant et les émissions de CO2 des pays africains en 2018. La forme de la relation est de type puissance avec un exposant de $1.41 > 1$ ce qui indique que l'accroissement des émissions n'est pas linéaire mais de plus en plus rapide lorsque le développement augmente. Un pays dont le revenu est de

5000 \$/hab. émettra moins de 1 tonne de CO₂ par habitant alors qu'un pays dont le revenu est de 10 000 \$/hab émettra plus de 2 tonnes et un pays dont le revenu est de 20 000 \$ par habitant plus de 6 tonnes !