Université Assane SECK de Ziguinchor
UFR des Sciences et Technologies (ST)
Département Informatique
Licence 2 : Ingénierie Informatique

Régression linéaire avec le logiciel R

Mor NDONGO
mndongo@univ-zig.sn

August 7, 2022

Sommaire

Sommaire

1. On considère deux attributs (ou caractères) des unités statistiques d'une population Ω.

1. On considère deux attributs (ou caractères) des unités statistiques d'une population Ω.

2. Ces deux attributs sont respectivement évalués sur des échelles de classification numériques et on note X et Y les variables qui expriment les évaluations de ces attributs sur les unités statistiques.

1. On considère deux attributs (ou caractères) des unités statistiques d'une population Ω.

2. Ces deux attributs sont respectivement évalués sur des échelles de classification numériques et on note X et Y les variables qui expriment les évaluations de ces attributs sur les unités statistiques.

3. L'attribut exprimé par Y est celui dont on veut étudier les variations d'une unité statistique à l'autre. L'attribut exprimé par X traduit une hétérogénéité de la population dont dépend les variations moyennes de Y .

Exemple

L'inhalation régulière du monoxyde de carbone étant considérée comme nuisible à la santé, une étude sur le tabagisme s'intéresse à la variation de la quantité de monoxyde de carbone d'une cigarette à l'autre.

Les données collectées à cet effet résultent de mesures effectuées pour l'évaluation de deux attributs :

Exemple

L'inhalation régulière du monoxyde de carbone étant considérée comme nuisible à la santé, une étude sur le tabagisme s'intéresse à la variation de la quantité de monoxyde de carbone d'une cigarette à l'autre.

Les données collectées à cet effet résultent de mesures effectuées pour l'évaluation de deux attributs :

1. le monoxyde de carbone produit par la combustion d'une cigarette et dont la quantité mesurée est exprimée en mg et noté par Y ;

Exemple

L'inhalation régulière du monoxyde de carbone étant considérée comme nuisible à la santé, une étude sur le tabagisme s'intéresse à la variation de la quantité de monoxyde de carbone d'une cigarette à l'autre.

Les données collectées à cet effet résultent de mesures effectuées pour l'évaluation de deux attributs :

1. le monoxyde de carbone produit par la combustion d'une cigarette et dont la quantité mesurée est exprimée en mg et noté par Y ;
2. le goudron contenu dans cette cigarette dont la quantité mesurée est exprimée en mg et notée par X.

Exemple

L'inhalation régulière du monoxyde de carbone étant considérée comme nuisible à la santé, une étude sur le tabagisme s'intéresse à la variation de la quantité de monoxyde de carbone d'une cigarette à l'autre.

Les données collectées à cet effet résultent de mesures effectuées pour l'évaluation de deux attributs :

1. le monoxyde de carbone produit par la combustion d'une cigarette et dont la quantité mesurée est exprimée en mg et noté par Y ;
2. le goudron contenu dans cette cigarette dont la quantité mesurée est exprimée en mg et notée par X.

Question : La quantité de goudron est-il un bon indicateur de la quantité moyenne de monoxyde de carbone émise par une cigarette ?

Variable explicative et variable réponse

- Y est appelée variable réponse, ou variable à expliquer ou variable dépendante ;
- X est appelée variable explicative, covariable ou variable indépendante.

Sommaire

Format usuel de présentation des données

Les données issues de l'évaluation des deux attributs sur les unités statistiques d'un échantillon Ω_n de taille n se présentent sous la forme $\{(x_i, y_i),\ i = 1 : n\}$.

Obs	Y	X
1	y_1	x_1
2	y_2	x_2
...	...	...
i	y_i	x_i
...	...	...
n	y_n	x_n

Présentation des données

Exemple

	Marque	Monoxide de carbonne (mg)	Goudron (mg)
1	Alpine	13.6	14.1
2	Benson&Hedges	16.6	16.0
3	BullDurham	23.5	29.8
4	CamelLights	10.2	8.0
5	Carlton	5.4	4.1
6	Chesterfield	15.0	15.0
7	GoldenLights	9.0	8.8
8	Kent	12.3	12.4
9	Kool	16.3	16.6
10	L&M	15.4	14.9
11	LarkLights	13.0	13.7
12	Marlboro	14.4	15.1
13	Merit	10.0	7.8
14	MultiFilter	10.2	11.4
15	NewportLights	9.5	9.0
16	Now	1.5	1.0
17	OldGold	18.5	17.0
18	PallMallLight	12.6	12.8
19	Raleigh	17.5	15.8
20	SalemUltra	4.9	4.5
21	Tareyton	15.9	14.5
22	True	8.5	7.3
23	ViceroyRichLight	10.6	8.6
24	VirginiaSlims	13.9	15.2
25	WinstonLights	14.9	12.0

Représentation graphique

Diagramme de dispersion

Le diagramme de dispersion (ou nuage de points) est la présentation graphique des données dans un repère d'axes orthogonaux telle que l'unité statistique ω_i de l'échantillon observé correspond au point de coordonnées $(\phi(x_i), y_i)$.

Représentation graphique

Diagramme de dispersion

Le diagramme de dispersion (ou nuage de points) est la présentation graphique des données dans un repère d'axes orthogonaux telle que l'unité statistique ω_i de l'échantillon observé correspond au point de coordonnées $(\phi(x_i), y_i)$.

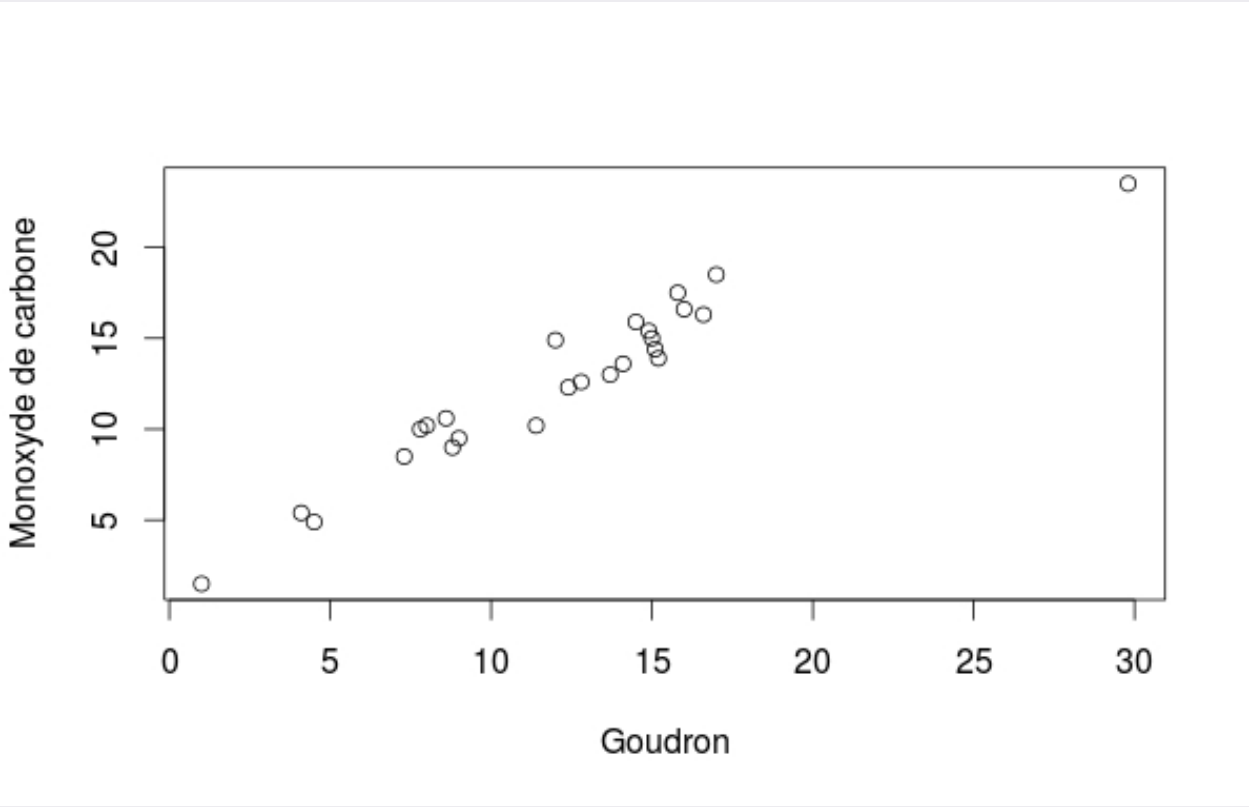
Mise en oeuvre sous R

Importer les données et tracer le diagramme de dispersion

```
> cigarettedata
<-read.table("/Bureau/courslogicielR/data/cigarettedata.csv",
header=TRUE, dec=",", quote="")
> X = cigarettedata[[3]]; Y = cigarettedata[[2]];
> plot(X,Y,xlab="Goudron",ylab="Monoxyde de carbone")
```

Représentation graphique

Exemple



Sommaire

Régression linéaire simple

Présentation du modèle

- On considère que pour toute valeur x de la variable explicative la réponse y observée peut s'écrire
  $y = a + bx + \varepsilon$

Régression linéaire simple

Présentation du modèle

- On considère que pour toute valeur x de la variable explicative la réponse y observée peut s'écrire
  $y = a + bx + \varepsilon$
- $a + bx$ est la valeur moyenne attendue de la réponse lorsque la condition d'hétérogénéité exprimée par X vaut x.

Régression linéaire simple

Présentation du modèle

- On considère que pour toute valeur x de la variable explicative la réponse y observée peut s'écrire
  $y = a + bx + \varepsilon$
- $a + bx$ est la valeur moyenne attendue de la réponse lorsque la condition d'hétérogénéité exprimée par X vaut x.
- ε est une valeur non observable et qui exprime la variabilité des réponses particulières y_i par rapport à la valeur attendue $a + bx$ qui correspond à la condition d'hétérogénéité x.

Régression linéaire simple

Présentation du modèle

- On considère que pour toute valeur x de la variable explicative la réponse y observée peut s'écrire
  $y = a + bx + \varepsilon$
- $a + bx$ est la valeur moyenne attendue de la réponse lorsque la condition d'hétérogénéité exprimée par X vaut x.
- ε est une valeur non observable et qui exprime la variabilité des réponses particulières y_i par rapport à la valeur attendue $a + bx$ qui correspond à la condition d'hétérogénéité x.
- On considère que la variabilité de la réponse par rapport à la valeur attendue a + bx est indépendante de x. Elle est évaluée par un paramètre σ^2 inconnu.

Objectif de l'ajustement du modèle aux données

- Le modèle

$$y = a + bx + \varepsilon$$

qui relie chaque réponse observée y_i de Y à la valeur x_i de la variable explicative X qui lui est associée dépend de 3 paramètres inconnus : a, b et σ^2.

Objectif de l'ajustement du modèle aux données

- Le modèle

$$y = a + bx + \varepsilon$$

qui relie chaque réponse observée y_i de Y à la valeur x_i de la variable explicative X qui lui est associée dépend de 3 paramètres inconnus : a, b et σ^2.

- Les paramètres a, b et σ^2 sont inconnus et l'objectif du traitement statistique est de les évaluer à partir des données bivariées $\{(x_i, y_i),\ i = 1 : n\}$.

Régression linéaire simple

Critère des moindres carrés ordinaires

- Le modèle qui relie les réponses observées à la valeur x de la variable explicative X qui leur est associée dépend de 3 paramètres inconnus : a, b et σ^2. Pour spécifier complètement le modèle il faudra évaluer les 3 paramètres inconnus à partir des données observées $\{(x_i, y_i),\ i = 1 : n\}$.

Régression linéaire simple

Critère des moindres carrés ordinaires

- Le modèle qui relie les réponses observées à la valeur x de la variable explicative X qui leur est associée dépend de 3 paramètres inconnus : a, b et σ^2. Pour spécifier complètement le modèle il faudra évaluer les 3 paramètres inconnus à partir des données observées $\{(x_i, y_i),\ i = 1 : n\}$.
- Les évaluations statistiques (estimations) des paramètres a et b sont obtenues à partir du critère des moindres carrés ordinaires

$$Q_n(a, b) = \sum_{i=1}^{n} [y_i - a - bx_i]^2$$

Régression linéaire simple

Critère des moindres carrés ordinaires

- Le modèle qui relie les réponses observées à la valeur x de la variable explicative X qui leur est associée dépend de 3 paramètres inconnus : a, b et σ^2. Pour spécifier complètement le modèle il faudra évaluer les 3 paramètres inconnus à partir des données observées $\{(x_i, y_i),\ i = 1 : n\}$.
- Les évaluations statistiques (estimations) des paramètres a et b sont obtenues à partir du critère des moindres carrés ordinaires

$$Q_n(a, b) = \sum_{i=1}^{n} [y_i - a - bx_i]^2$$

- Les paramètres a et b sont estimés par les valeurs $\widehat{a}_n$ et $\widehat{b}_n$ qui réalisent le minimum de Q_n.

Régression linéaire simple

Minimisation du critère des moindres carrés ordinaires

- Soit

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

- Les solution du problème sont :

$$\widehat{a}_n = \bar{y} - \widehat{b}_n \bar{x} \qquad \widehat{b}_n = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$\widehat{\sigma}_n^2 = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \widehat{a}_n - \widehat{b}_n x_i)^2$$

Définition (Réponses ajustées)

On appelle valeurs ajustées par le modèle les valeurs $\widehat{y}_i$:

$$\widehat{y}_i = \widehat{a}_n + \widehat{b}_n x_i$$

Définition (valeurs résiduelles)

On appelle valeurs résiduelles les écarts $\widehat{\varepsilon}_i$:

$$\widehat{\varepsilon}_i = y_i - \widehat{y}_i$$

Analyse graphique des résidus

On représente les observations dans l'espace rapporté à un système d'axes orthogonaux par les points de coordonnées $(x_i, \varepsilon_i),\ i = 1 : n$.

Les données sont jugés compatibles avec l'hypothèse de linéarité si les points représentatifs des observations ne présentent pas une structure évidente suivant une relation fonctionnelle entre abscisses et ordonnées.

Coefficient de détermination

On appelle coefficient de détermination le rapport de la variabilité expliquée par la régression : $\sum_{i=1}^{n}(\widehat{y}_i - \bar{y}_n)^2$ sur la variabilité totale des y_i : $\sum_{i=1}^{n}(y_i - \bar{y}_n)^2$

$$R_n^2 = \frac{\sum_{i=1}^{n}(\widehat{y}_i - \bar{y}_n)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_n)^2}$$

$$\bar{y}_n = \frac{1}{n}\sum_{i=1}^{n}\widehat{y}_i$$

Interprétation du coefficient de détermination

Fort logiquement, le R_n^2 prend ses valeur dans $[0,\ 1]$: au pire, le modèle n'explique rien, au mieux il explique 100% de la variance de Y.

Si pour un modèle, on trouve $R_n^2 = 0.98$, on dira que 98% de la variance est due à la régression ou encore que la variance résiduelle représente 2% de la variance des observations y_i.

Mise en oeuvre sous R

```
> tabacdata=read.table("cigarettedata.csv", header=TRUE)
> attach(tabacdata)
> X = goudron; Y=monoxyde
> result = lm(Y ~ X)
> summary(result)
```

Mise en oeuvre sous R

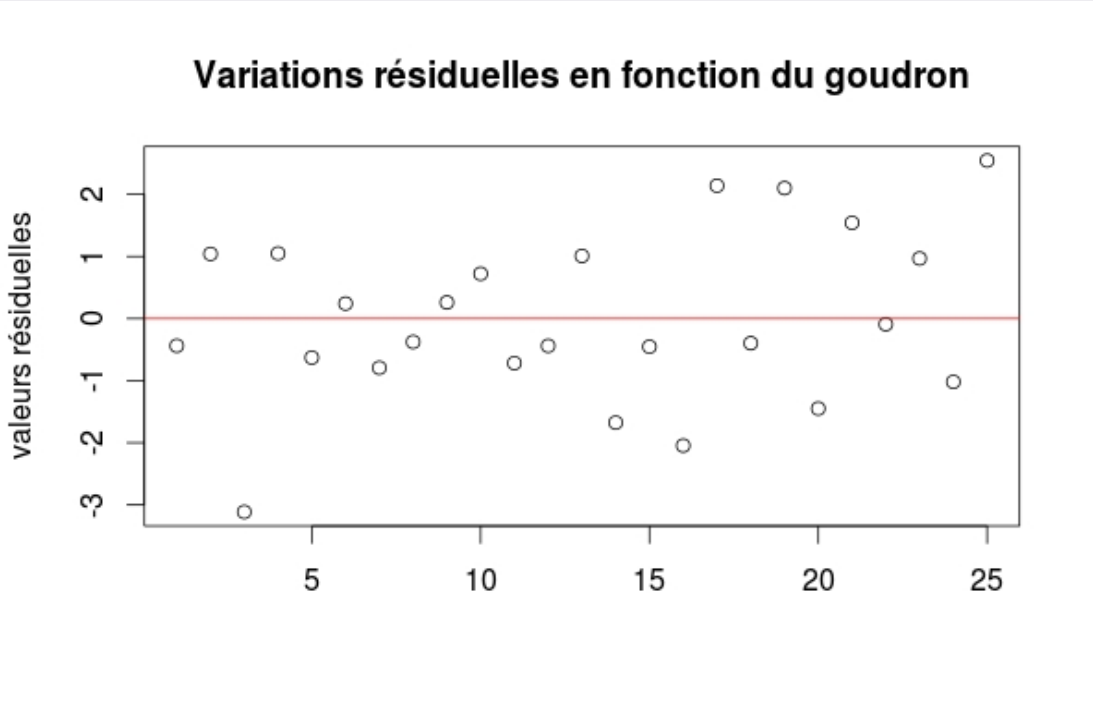
```
> tabacdata=read.table("cigarettedata.csv", header=TRUE)
> attach(tabacdata)
> X = goudron; Y=monoxyde
> result = lm(Y ~ X)
> summary(result)
```

Résultats

$\widehat{a}_n = 2.74328$ $\widehat{b}_n = 0.80098$ $\widehat{\sigma}^2 = 1.951609$
$R^2 = 0.9168$

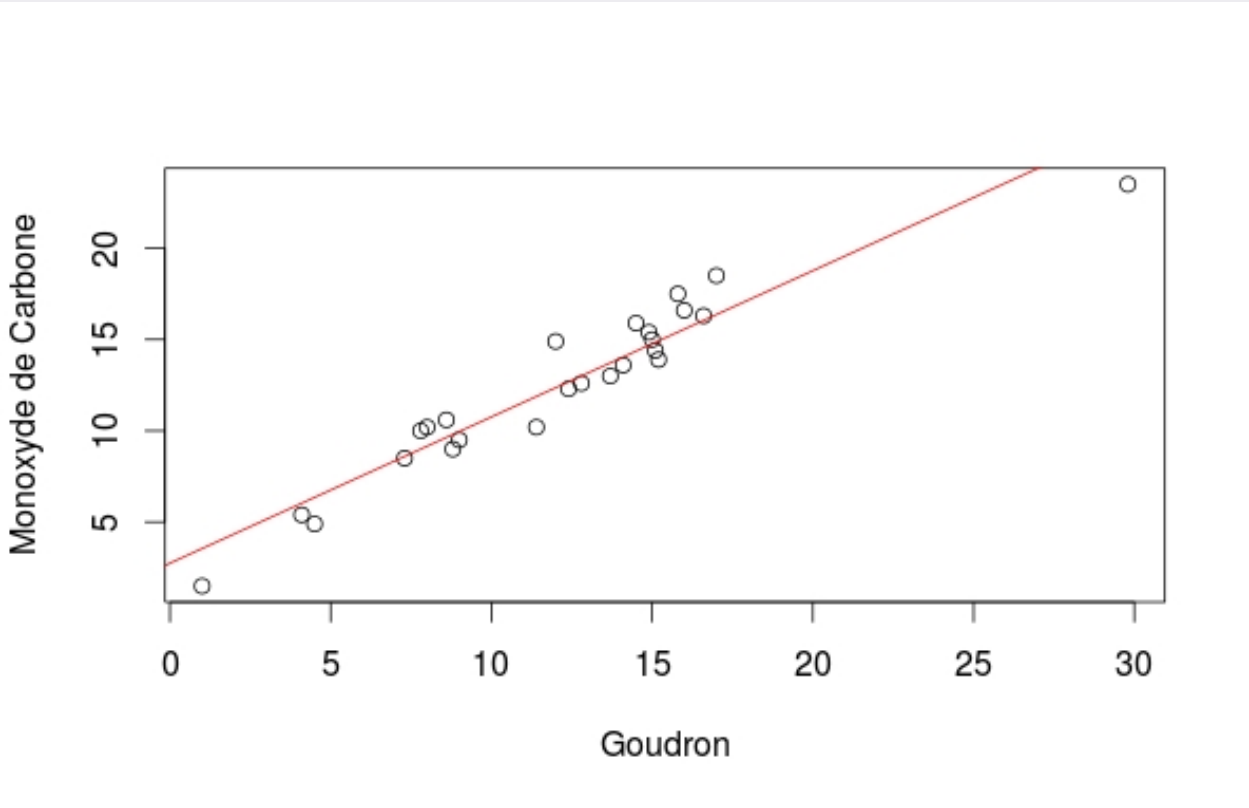
Mise en oeuvre sous R

```
> resid = result$residuals
> plot(resid,ylab="valeurs résiduelles",xlab="",main="Variations résiduelles en fonction du goudron")
> abline(h=0,col="red")
```



Mise en oeuvre sous R

```
> plot(X,Y,xlab="Goudron",ylab="Monoxyde de Carbone")
> abline(2.74328,0.80098,col="red")
```



Sommaire

Régression linéaire multiple

Objectifs

Technique de modélisation qui permet de mettre en équation une relation entre une variable endogène (à expliquer) et n variables exogènes (explicatives).

Cette technique est couramment utilisée lorsque l'on souhaite prédire la réalisation d'une variable de type continue (intervalle ou ratio) à l'aide d'un ensemble de variables, dits prédicteurs, du même type; des prédicteurs de type catégoriels pouvant aussi être considérés.

Régression linéaire multiple

Modèle général de régression linéaire

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots \beta_p X_{pi} + \varepsilon_i$$

- Les coefficients $\beta_0, \beta_1, \ldots, \beta_p$ sont les paramètres inconnus du modèle.
- Y est la variable expliquée (endogène)
- X_i sont les variables explicatives (exogènes)
- ε est le terme d'erreur

Hypothèses requises

- H_1 : $\mathbb{E}(\varepsilon_i) = 0,\ \forall i$.
- H_2 : $Var(\varepsilon_i) = \sigma_\varepsilon^2,\ \forall i$ (homoscédasticité).
- H_3 : $cov(\varepsilon_i, \varepsilon_j) = 0;\ \forall i \neq j$ (non autocorrélation des résidus)
- H_4 : $cov(\varepsilon_i, X_i) = 0;\ \forall i$ (exogénéité)
- H_5 : ε_i suit une loi normale, c'est-à-dire $\varepsilon_i \sim \mathcal{N}(0,\ \sigma_\varepsilon^2)$.

Régression linéaire multiple

Exemple : Etude de cas

On veut expliquer le produit intérieur brut (PIB) en fonction de 6 facteurs explicatifs :

- formation brut de capital (FBC)
- dépense nationale brut (BND)
- dépense de consommation finale des ménages (DCFM)
- dépenses de consommation finale des administrations publiques (DCFAP)
- la valeur ajoutée de l'agriculture (AVA)
- le revenu intérieur brut (RIB)

- **Les données sont collectées 1960 à 2014.**

Régression linéaire multiple

Méthodologie

- Identification de la liaison de type linéaire entre Y et les X_i
- Estimation de l'équation de régression (des coefficients)
- Validation du modèle (Analyse des résidus)
- Qualité du modèle de régression linéaire
- Prévision

Représentation graphique

On peut représenter le PIB avec chacune des variables explicatives (FBC, BND, DCFM, DCFAP, AVA, RIB). Pour cela on utilise la fonction splom du package lattice.

Mise en oeuvre sous R

```
> basePIB = read.csv2("basePIB.csv")
> DATA = data.frame(PIBt,FBCt,DNBt,DCFMt,DCFAPt,AVAt,RIBt)
> library(lattice)
> splom(DATA)
```

On obtient une matrice de nuage de points. La dernière ligne de la matrice donne les nuages de points entre la variable PIB et les variables explicatives.

Mise en oeuvre sous R

Matrice de nuages de points

Au vue de ce diagramme de dispersion, nous pouvons dire que la relation entre la variable PIB et toutes les autres variables est linéaire.

Matrice de corrélation

L'examen de la matrice du nuage de points montre que la variable expliquée (PIB) est linéairement corrélée avec chacune des variables explicatives. Nous pouvons ensuite calculer la matrice de corrélation pour apprécier l'intensité de cette corrélation. Ceci s'obtient sous R par la commande :

```
> cor(DATA
```

Les variables explicatives sont fortement corrélées (positivement) avec la variable à expliquer. La corrélation est un bon indicateur de mesure entre les variables.

Estimation du modèle

$$PIB = \beta_0 + \beta_1 FBC + \beta_2 BND + \beta_3 DCFM + \beta_4 DCFAP + \beta_5 AVA + \beta_6 RIB$$

```
> reg = lm(PIBt~ FBCt+DNBt + DCFMt + DCFAPt + AVAt + RIBt)
> summary(reg)
```

Régression linéaire multiple
Estimation des coefficients

Résultats

```
Call:
lm(formula = PIBt ~ 1 + FBCt + DNBt + DCFMt + DCFAPt + AVAt +
    RIBt)

Residuals:
       Min          1Q      Median          3Q         Max
-1.392e+11 -2.862e+10  8.495e+09  3.826e+10  1.186e+11

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -7.305e+10  8.478e+10  -0.862 0.393172
FBCt        -6.950e-01  2.947e-01  -2.359 0.022463 *
DNBt         1.046e+00  2.754e-01   3.797 0.000412 ***
DCFMt       -2.625e-01  2.540e-01  -1.034 0.306534
DCFAPt      -6.579e-01  3.760e-01  -1.750 0.086548 .
AVAt         5.799e-01  1.888e-01   3.072 0.003498 **
RIBt         2.143e-01  1.380e-01   1.553 0.127041
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.506e+10 on 48 degrees of freedom
Multiple R-squared:  0.9982,    Adjusted R-squared:  0.9979
F-statistic:  4376 on 6 and 48 DF,  p-value: < 2.2e-16
```

Estimation du modèle

L'équation du modèle (estimé) de régression linéaire est donnée :

$$\begin{aligned} PIB &= -7.305.10^{10} - 0.6950 \times FBC + 1.046 \times BND \\ &- 0.2625 \times DCFM - 0.6579 \times DCFAP + 0.5799 \times AVA \\ &+ 0.2143 \times RIB \end{aligned}$$

Significativité globale

L'appréciation de la qualité globale du modèle se fait avec la statistique de Fisher, qui montre si les variables explicatives ont une influence sur la variable dépendante.
Les hypothèses du teste sont :
H_0 : $\beta_0 = \beta_1 = \cdots = \beta_6 = 0$ (le modèle n'est pas globalement significatif)
H_1 : il existe au moins un $\beta_i \neq 0$ (le modèle est globalement significatif)

```
Residual standard error: 5.506e+10 on 48 degrees of freedom
Multiple R-squared:  0.9982,    Adjusted R-squared:  0.9979
F-statistic:  4376 on 6 and 48 DF,  p-value: < 2.2e-16
```

P-value : $< 2.2e^{-16} < 5\%$ donc le modèle est globalement significatif au seuil de 5%.

Teste de significativité individuelle

Ce test permet de voir parmi les variables explicatives (exogènes) utilisées dans le modèle, celles qui ont une influence significative sur la variable expliquée (endogène).

Les hypothèses testées sont :

H_0 : $\beta_i = 0$ (effet non significatif)

H_1 : $\beta_i \neq 0$ (effet significatif)

Teste de significativité individuelle

```
Call:
lm(formula = PIBt ~ 1 + FBCt + DNBt + DCFMt + DCFAPt + AVAt +
    RIBt)

Residuals:
       Min          1Q      Median          3Q         Max
-1.392e+11 -2.862e+10  8.495e+09  3.826e+10  1.186e+11

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -7.305e+10  8.478e+10  -0.862 0.393172
FBCt        -6.950e-01  2.947e-01  -2.359 0.022463 *
DNBt         1.046e+00  2.754e-01   3.797 0.000412 ***
DCFMt       -2.625e-01  2.540e-01  -1.034 0.306534
DCFAPt      -6.579e-01  3.760e-01  -1.750 0.086548 .
AVAt         5.799e-01  1.888e-01   3.072 0.003498 **
RIBt         2.143e-01  1.380e-01   1.553 0.127041
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.506e+10 on 48 degrees of freedom
Multiple R-squared:  0.9982,    Adjusted R-squared:  0.9979
F-statistic:  4376 on 6 and 48 DF,  p-value: < 2.2e-16
```

Seules trois variables sont significatives au seuil de 5% : la formation brute de capital, la dépense nationale brute et la valeur ajoutée de l'agriculture.

Le pourcentage de variance expliquée

Utilisé dans de nombreuses analyses statistiques comme critère de qualité d'ajustement d'un modèle.
Le coefficient de détermination ou R^2 (appelé R-deux) correspond non seulement au carré du coefficient de corrélation multiple entre Y et $X_1, \ldots, X_p$. C'est également le carré de la corrélation entre la variable observée, Y, et la valeur prédite, $\hat{Y}$ obtenue à partir du modèle de régression.

$$R^2 = \frac{V(\hat{Y})}{V(Y)}$$

```
Residual standard error: 5.506e+10 on 48 degrees of freedom
Multiple R-squared:  0.9982,    Adjusted R-squared:  0.9979
F-statistic:  4376 on 6 and 48 DF,  p-value: < 2.2e-16
```

Par conséquent, 99.82% de variance est expliqué par notre modèle.

R^2 Ajusté

Le R-deux Ajusté est davantage utilisé que le R^2 car il ne dépend pas du nombre de variables.

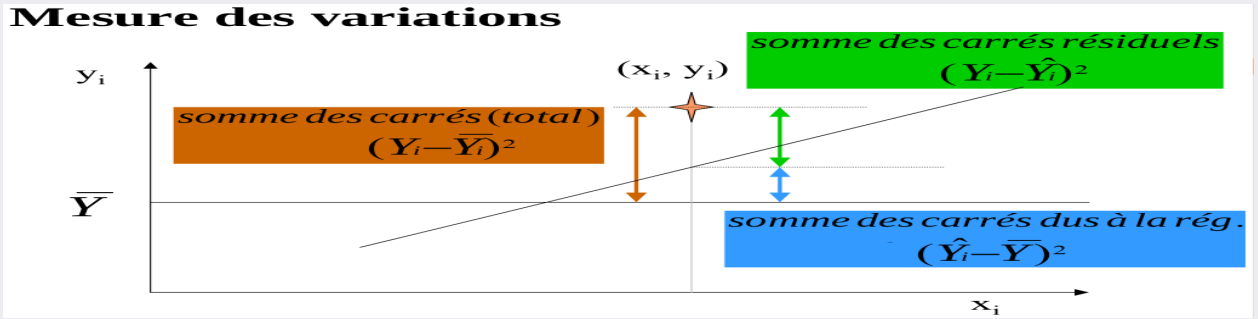
$$\overline{R}^2 = R^2 - \frac{p(1-R^2)}{N-p-1}$$

où p est le nombre de variables indépendantes et N le nombre d'observations Récapitulatif.

```
Residual standard error: 5.506e+10 on 48 degrees of freedom
Multiple R-squared:  0.9982,    Adjusted R-squared:  0.9979
F-statistic:  4376 on 6 and 48 DF,  p-value: < 2.2e-16
```

Par conséquent, 99.79% de variance est expliqué par notre modèle.

Analyse de la variance



$$\sum_{i=1}^{N}\left(Y_i-\overline{Y}\right)^2=\sum_{i=1}^{N}\left(Y_i-\hat{Y}_i\right)^2+\sum_{i=1}^{N}\left(\hat{Y}_i-\overline{Y}\right)^2$$

la somme des carrés des écarts à la moyenne (variance empirique de Y) est égale à la somme des carrés des résidus (variance empirique des résidus) plus la variance empirique de $\hat{Y}$ (variance empirique de $\hat{Y}$).

Hypothèses de la regression linéaire

Afin d'avoir des tests fiables, il est indispensable de vérifier certaines hypothèses fondamentales de base.

- Hypothèse 1 : Normalité de Y
- Hypothèse 2 : Homocédasticité des erreurs
- Hypothèse 3 : Les résidus doivent être indépendants, Normaux, centrés et non corrélées avec les variables explicatives.

Hypothèse 1 : Normalité de Y

La régression linéaire sous-entend que Y est normalement distribuée. Afin de vérifier ce type d'hypothèse, R vous propose plusieurs outils :

1. test de Kolmogorov-Smirnov (ks.test),
2. test de Shapiro-Wilk (shapiro.test),
3. test de Jarque-Bera (jarqueberaTest)

Graphiquement, vous disposez également de plusieurs outils permettant d'illustrer les tests proposés ci-dessus.

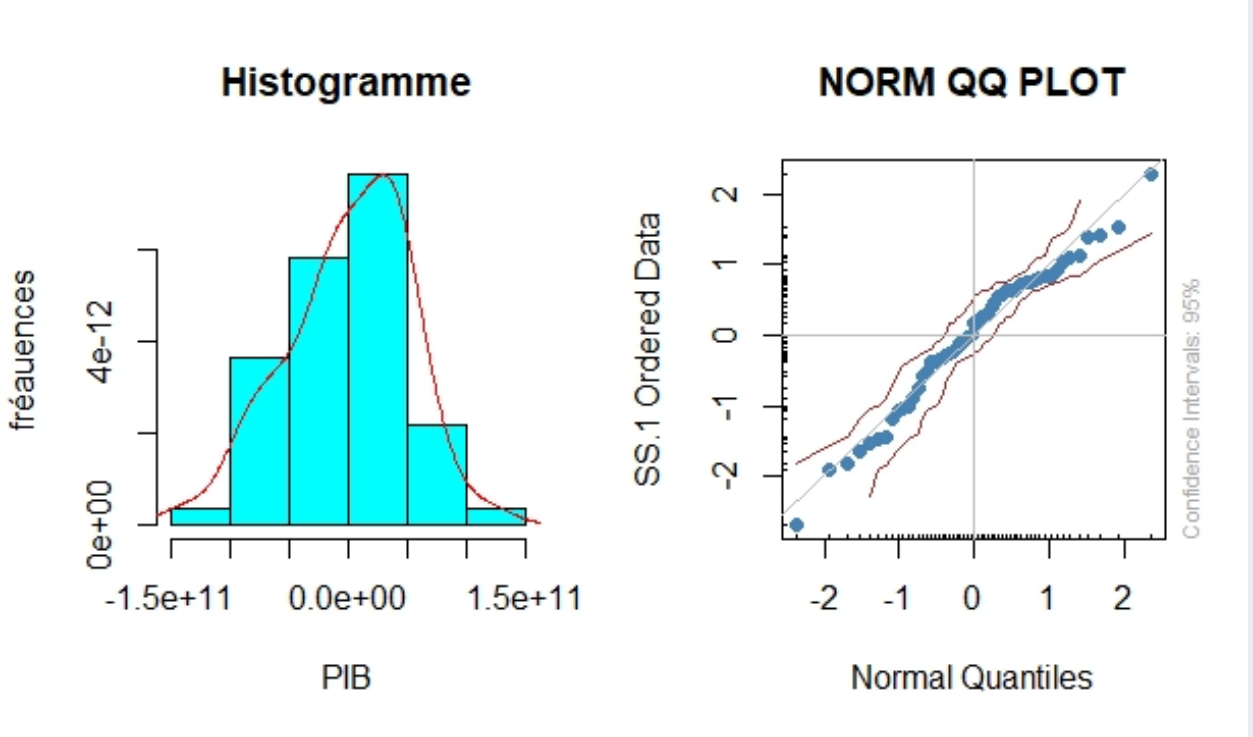
1. Histogramme (hist)
2. Diagramme Quantile-Quantile Normal (qqnorm)

Hypothèse 1 : Normalité de Y

```
> y = density(resid)
> hist(resid,probability = T,col=5,xlab = "PIB",ylab =
"fréauences",main="Histogramme")
> lines(y,col=2)
> qqnorm(resid)
> qqline(resid,col=2)
```

On peut aussi utiliser la fonction qqnormPlot() du package fBasics qui donne une représentation des quantiles avec l'intervalle de confiance.

Résultats


Histogramme
fréauences
4e-12
0e+00
-1.5e+11
0.0e+00
1.5e+11
PIB
NORM QQ PLOT
SS.1 Ordered Data
2
1
0
-1
-2
-2
-1
0
1
2
Normal Quantiles
Confidence Intervals: 95%

Hypothèse 1 : Normalité de Y

H_0 : les résidus sont distribués suivant une loi normale
H_1 : La distribution n'est pas normale

```
> shapiro.test(resid)
```

$p - value = 0.3927 < 5\%$, donc on accepte H_0. D'où avec un risque de 5%, les résidus sont distribués suivant une loi normale.

Hypothèse 2 : Homocédasticité des erreurs

En régression, les vrais erreurs ε_i sont supposés être indépendants de moyenne 0 et de variance constante σ^2. Si le modèle est approprié pour les données, les résidus observés devraient avoir un comportement similaire.

Hypothèse 2 : Homocédasticité des erreurs

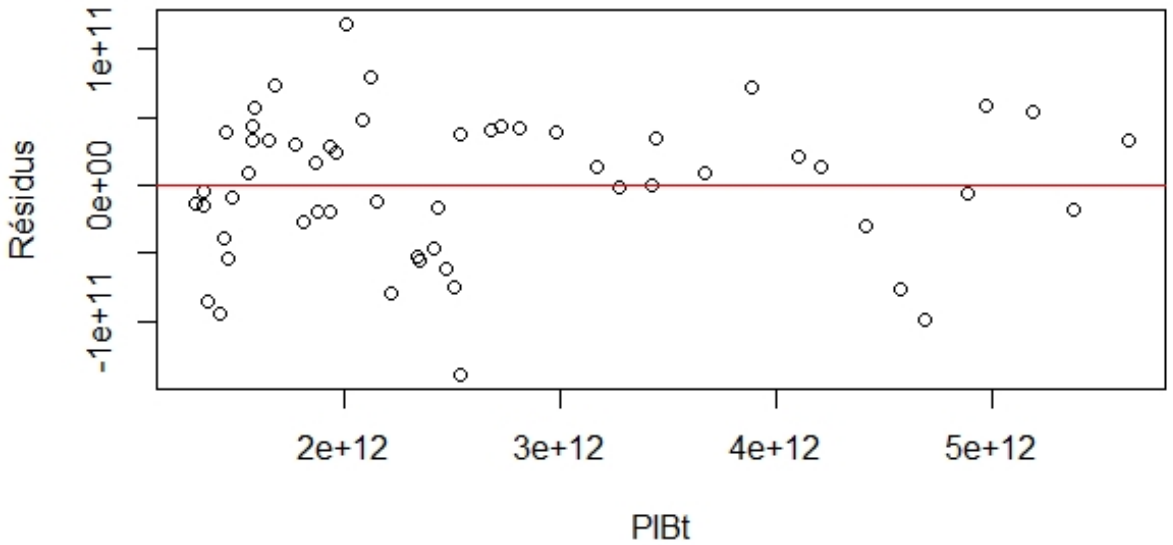
Variance constante : Cette étude est essentiellement graphique. On utilise :

- le graphe des résidus fonction des X
- Le graphe (pred, res).

Si l'hypothèse de linéarité et d'homogénéité de variance sont vérifiées:

- il ne devrait pas y avoir de relation entre **pred** et **res**,
- et les résidus devraient se comporter de manière aléatoire le long d'une bande autour de 0.
- la variabilité des résidus n'augmente pas en fonction de l'ampleur des valeurs prévues.

Hypothèse 2 : Homocédasticité des erreurs



Hypothèse 3 : Absence d'autocorrélation des résidus

Le test de Durbin-Watson est un test d'absence d'autocrrélation d'ordre 1 sur le résidu de régression linéaire. Il teste l'hypothèse $H_0 \ : \ \rho = 0$ (non autocorrélation). La statistique du test est :

$$DW = \frac{\sum_{t=2}^{T}(\hat{\varepsilon}_t - \hat{\varepsilon}_{t-1})^2}{\sum_{t=2}^{T}\hat{\varepsilon}_t^2} \simeq 2(1-\hat{\rho}) \in (0,4)$$

où $\hat{\rho} = \sum_{t=2}^{T}\hat{\varepsilon}_{t-1}\hat{\varepsilon}_t / \sum_{t=2}^{T}\hat{\varepsilon}_t^2$.

- Le test est programmé dans dwtest() de lmtest et dur.waston() de car.
- Pratiquement une statistique $DW \ll 2$ peut être le signe d'une mauvaise spécification du modèle (par exemple, ajustement d'une tendance linéaire alors que la tendance réelle est quadratique).

Exemple : PIB réel : $5.18470e+12$

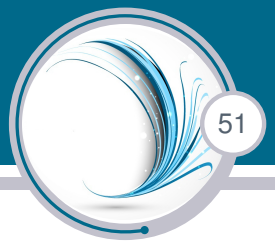
Prédir le PIB connaissant :

FBC	DNB	DCFM	DCFAP
$1.03068e+12$	$5.71187e+12$	$4.03214e+12$	$6.49050e+11$

AVA	RIB
$6.60358e+11$	$4.90142e+1$

```
> newx = data.frame(FBCt=1.03068e+12,DNBt=5.71187e+12,
DCFMt=4.03214e+12,DCFAPt=6.49050e+11,AVAt=6.60358e+11,
RIBt=4.90142e+12)
> AX = predict(reg,newdata = newx,se.fit = TRUE)
> AX$fit
```

La valeur prédite du PIB est : $PIB = 5.131213e+12$



L'objectif de la sélection de variables est de trouver un sous-ensemble de variables exogènes pertinentes et non-redondantes pour expliquer l'endogène. A cet effet, le logiciel R calcule l'AIC (Akaike Information Criterion) d'un modèle et faire ainsi des comparaisons. On peut ainsi tester tous les effets et garder le modèle "minimal" qui offre le meilleur compromis entre "ajustement aux données" et "nombre de paramètres". La fonction step permet de sélectionner automatiquement un tel modèle.

```
> modelFinal = step(reg)
> summary(modelFinal)
```

Le modèle finale est le modèle régressé après sélection des variables (les variables non sélectionnées ne seront pas considérées).

MERCI