Régression linéaire avec le logiciel R

Université Assane SECK de Ziguinchor

UFR des Sciences et Technologies

Département Informatique

Licence 2 : Ingénierie Informatique

Sommaire

Sommaire

Présentation du contexe et de la problématique

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 .

Présentation du contexe et de la problématique

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 .

Présentation du contexe et de la problématique

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 .

Présentation du contexte et de la problématique

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 ?

Présentation du contexte et de la problématique

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 ?

Présentation du contexte et de la problématique

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 ?

Présentation du contexte et de la problématique

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 ?

Présentation du contexte et de la problématique

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

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

Diagramme de dispersion

Définition

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 $(\phi(x_i), y_i)$.

Exemple 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")
```

Diagramme de dispersion

Définition

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 $(\phi(x_i), y_i)$.

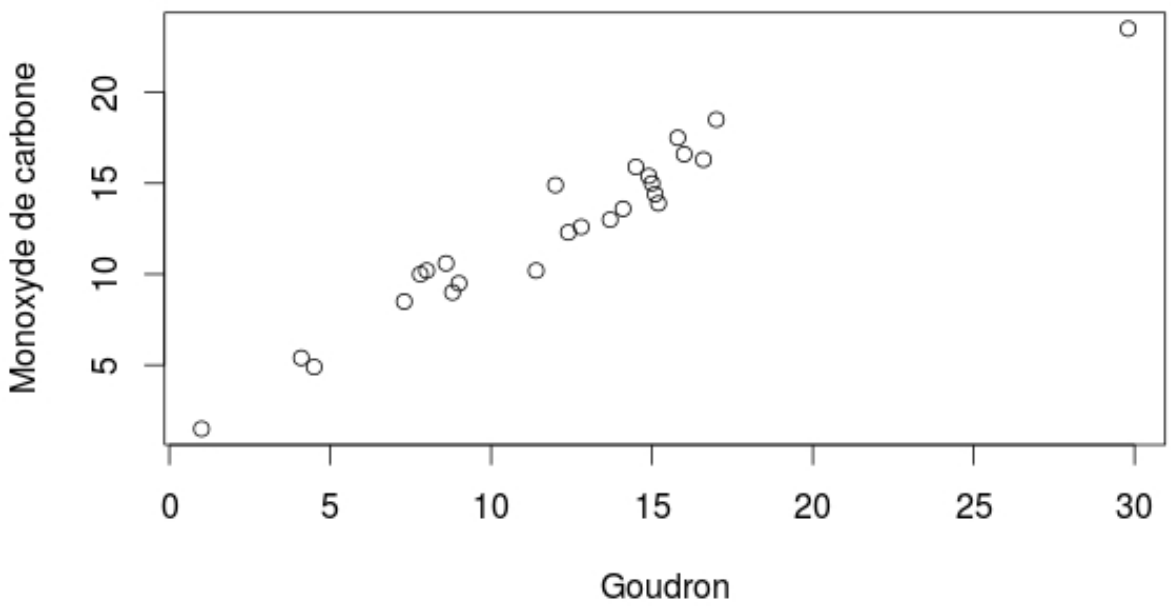
Exemple sous R

☞ 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")
```

Diagramme de dispersion

Exemple



Sommaire

Formulation du modèle linéaire simple

- 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.

Formulation du modèle linéaire simple

- 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.

Formulation du modèle linéaire simple

- 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.

Formulation du modèle linéaire simple

- 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.

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

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 $\{(x_i, y_i),\ i = 1 : n\}$.

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.

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.

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.

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 :

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

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

Eléments de diagnostic sur la validité du modèle

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 $\widehat{\varepsilon}_i$:

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

Eléments de diagnostic sur la validité du modèle

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 $\widehat{\varepsilon}_i$:

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

Diagnostic sur la linéarité : 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.

adéquation du modèle aux données

Coefficient de détermination

On appelle coefficient de détermination le quotient

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

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

adéquation du modèle aux données

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.

Exemple du tabagisme

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

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

Exemple du tabagisme

Mise en oeuvre sous R

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

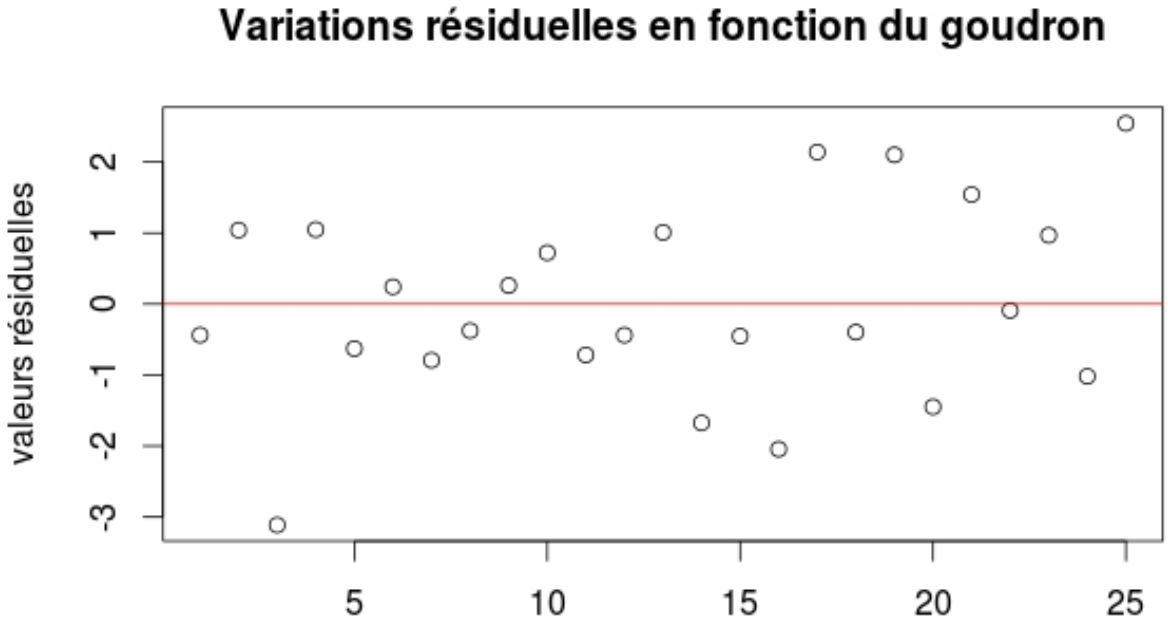
Résultats

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

Exemple du tabagisme

Mise en oeuvre sous R

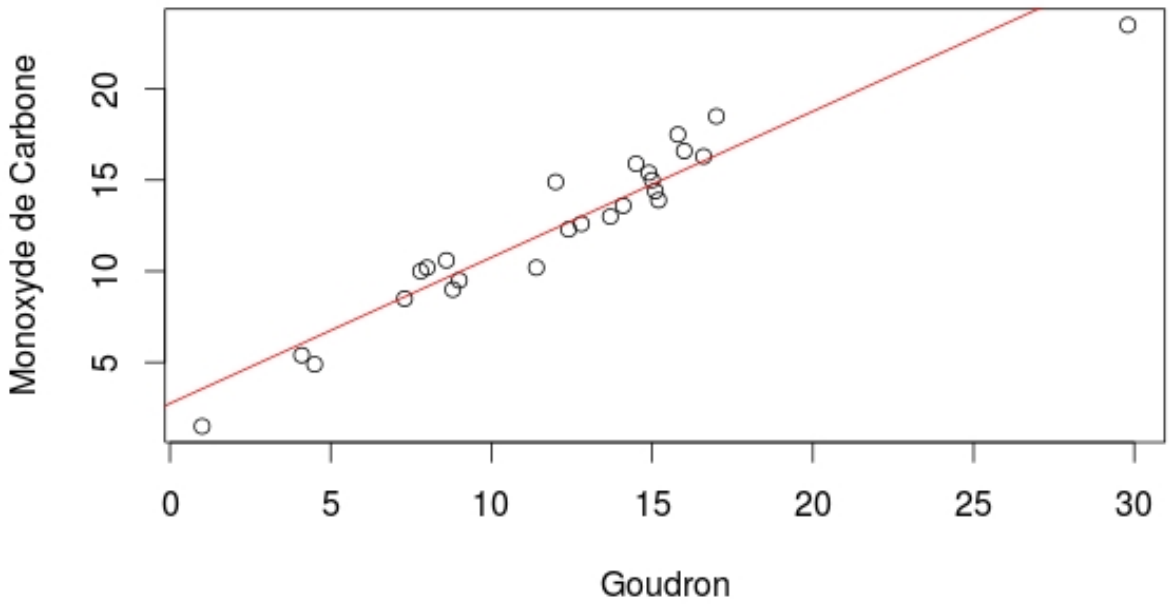
```
> resid = result$residuals
> plot(resid,ylab="valeurs résiduelles",xlab="",main="Variations
résiduelles en fonction du goudron")
> abline(h=0,col="red")
```



Exemple du tabagisme

Mise en oeuvre sous R

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



Sommaire