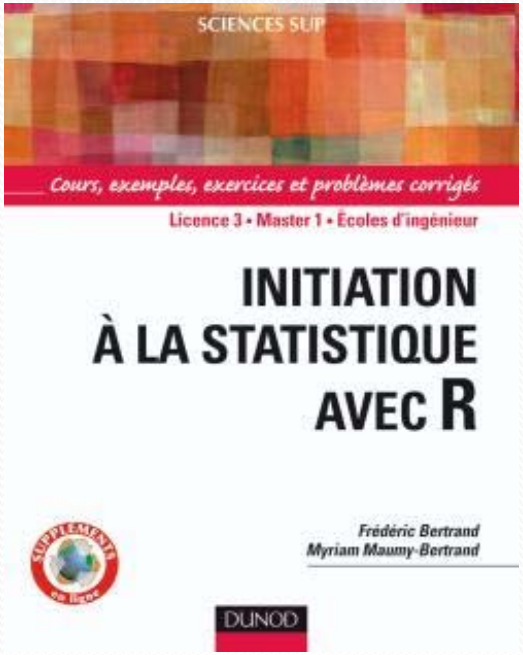
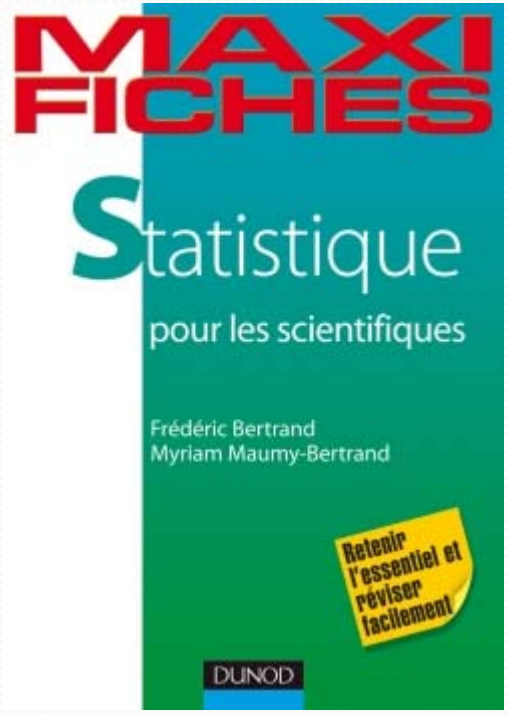
Régression linéaire simple

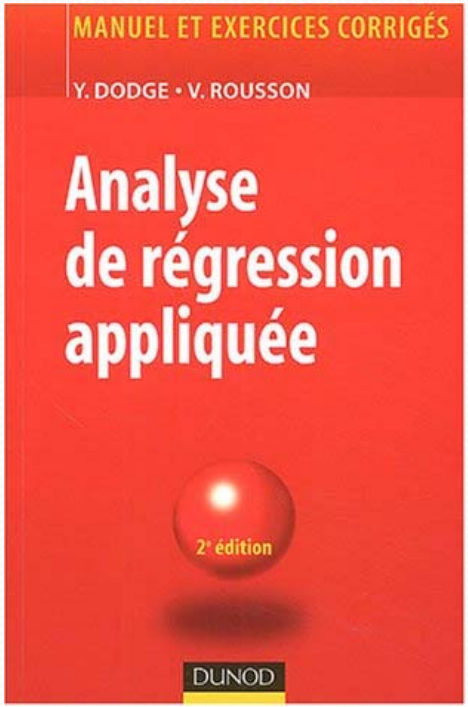
Myriam Maumy-Bertrand

MCB2A Formation continue 2014/2015

Références


SCIENCES SUP
Cours, exemples, exercices et problèmes corrigés
Licence 3 • Master 1 • Écoles d'ingénieur
INITIATION
À LA STATISTIQUE
AVEC R
Frédéric Bertrand
Myriam Maumy-Bertrand
DUNOD


MAXI
FICHES
Statistique
pour les scientifiques
Frédéric Bertrand
Myriam Maumy-Bertrand
Retenir l'essentiel et réviser facilement
DUNOD


MANUEL ET EXERCICES CORRIGÉS
Y. DODGE • V. ROUSSON
Analyse
de régression
appliquée
2e édition
DUNOD

Introduction

But : rechercher une relation stochastique qui lie deux ou plusieurs variables

Domaines :

- Physique, chimie, astronomie
- Biologie, médecine
- Géographie
- Economie
- …

1. Relation entre deux variables

Considérons X et Y deux variables.

Exemple : la taille (X) et la masse (Y)

But : savoir comment Y varie en fonction de X

Dans la pratique :

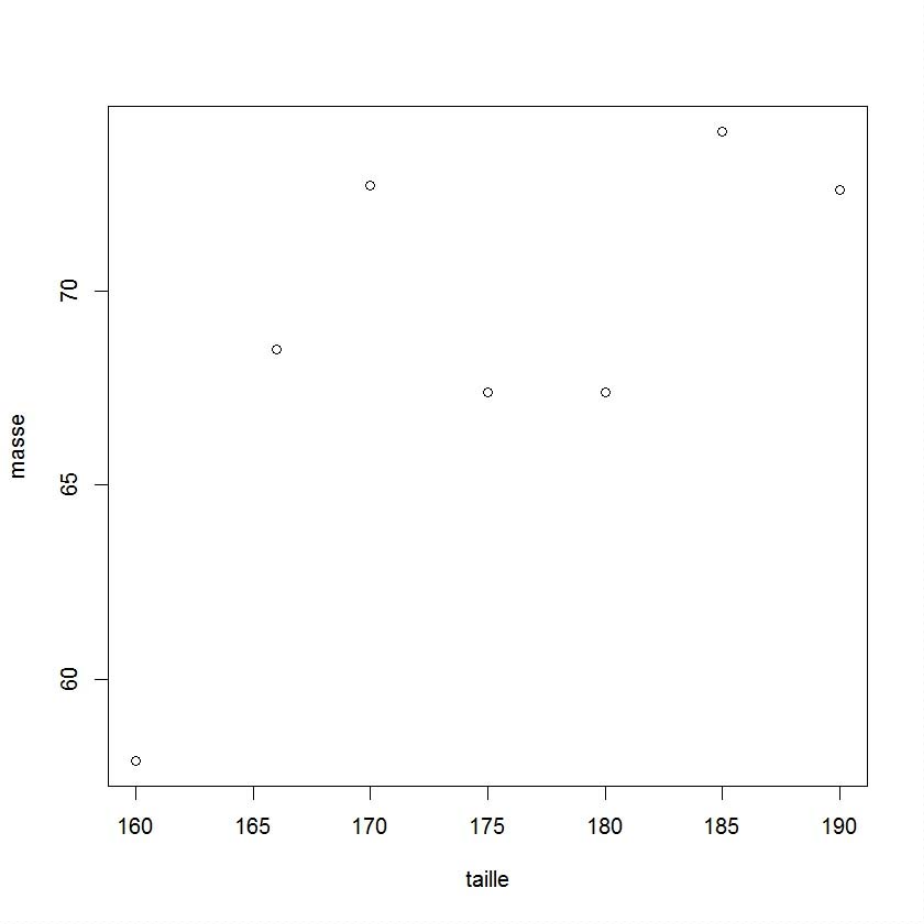
- Échantillon de n individus
- Relevé de la taille et de la masse pour l'individu i

⇒ Tableau d'observations ou données pairées.

1. Relation entre deux variables

Observations	Taille	Masse
1	160	57,9
2	165	68,5
3	170	72,7
4	175	67,4
5	180	67,4
6	185	74,1
7	190	72,6

1. Relation entre deux variables



1. Relation entre deux variables

Pour montrer qu'il existe une relation linéaire entre deux variables quantitatives, nous pouvons calculer le coefficient de corrélation linéaire de Bravais-Pearson. Pour une définition de ce coefficient, nous renvoyons à un des ouvrages recommandés. Avec **R**, nous obtenons :

```
> cor(masse,taille)
[1] 0.7128531
```

Comme ce coefficient est relativement proche de 1, il semblerait qu'il existe une relation linéaire entre ces deux variables.

2. Relation déterministe

Dans certains cas, la relation est exacte.

Exemples :

- X en euros, Y en dollars
- X distance ferroviaire, Y prix du billet.

$$Y = f(X)$$

où f est une fonction déterminée.

Exemples pour f : fonctions linéaires, fonctions affines…

2. Relation déterministe

Remarque importante :

Nous utiliserons le terme de fonction « linéaire » pour désigner une fonction « affine »

$$f(X) = \beta_0 + \beta_1 X$$

où β_0 et β_1 sont des réels fixés.

2. Relation déterministe

Exemple : X en Celsius, Y en Farenheit

$$Y=32 + 9/5\ X.$$

Ici nous avons en identifiant : $\beta_0 = 32$ et $\beta_1 = 9/5$.

Souvent nous savons que la relation entre X et Y est linéaire mais les coefficients sont inconnus.

2. Relation déterministe

En pratique comment faisons-nous ?

- Échantillon de *n* données
- Vérifier que les données sont alignées.

Si ce cas est vérifié, alors nous avons : un **modèle linéaire déterministe.**

2. Relation déterministe

Si ce cas n'est pas vérifié, alors nous allons chercher : **la droite qui ajuste le mieux l'échantillon**, c'est-à-dire nous allons chercher un **modèle linéaire non déterministe**.

Les *n* observations vont permettre de vérifier si la droite candidate est adéquate.

3. Relation stochastique

La plupart des cas ne sont pas des modèles linéaires déterministes !
(la relation entre *X* et *Y* n'est pas exacte)

Exemple : *X* la taille et *Y* la masse.
A 180 cm peuvent correspondre plusieurs masses :
75 kg, 85 kg, …
Les données ne sont plus alignées.
Pour deux masses identiques, nous avons deux tailles différentes.

3. Relation stochastique

Une hypothèse raisonnable : *X* et *Y* sont liés

Dans l'exemple précédent : plus un individu est grand, plus il est lourd

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

ε : est une variable qui représente le comportement individuel.

3. Relation stochastique

Exemple :

70 individus qui sont répartis de la façon suivante :

- 10 individus/taille
- 7 tailles (de 160 à 190 cm, pas de 5 cm)

Nous allons traiter cet exemple avec le logiciel **R**. Les lignes de commande qui permettent de tracer le graphe représentatif des couples sont présentées dans deux transparents.

3. Relation stochastique

Voici le tableau de données

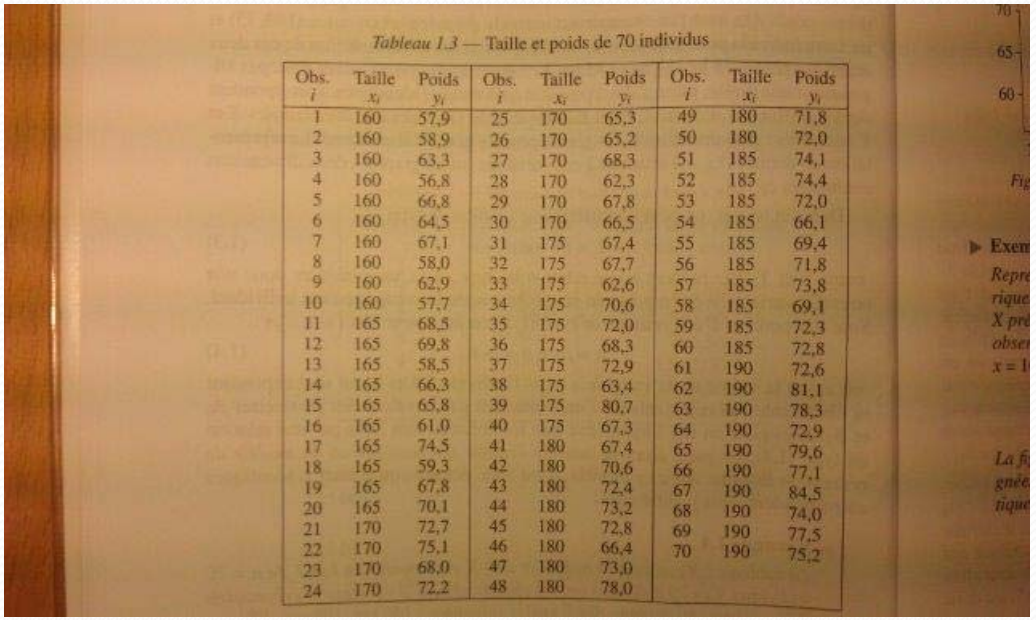


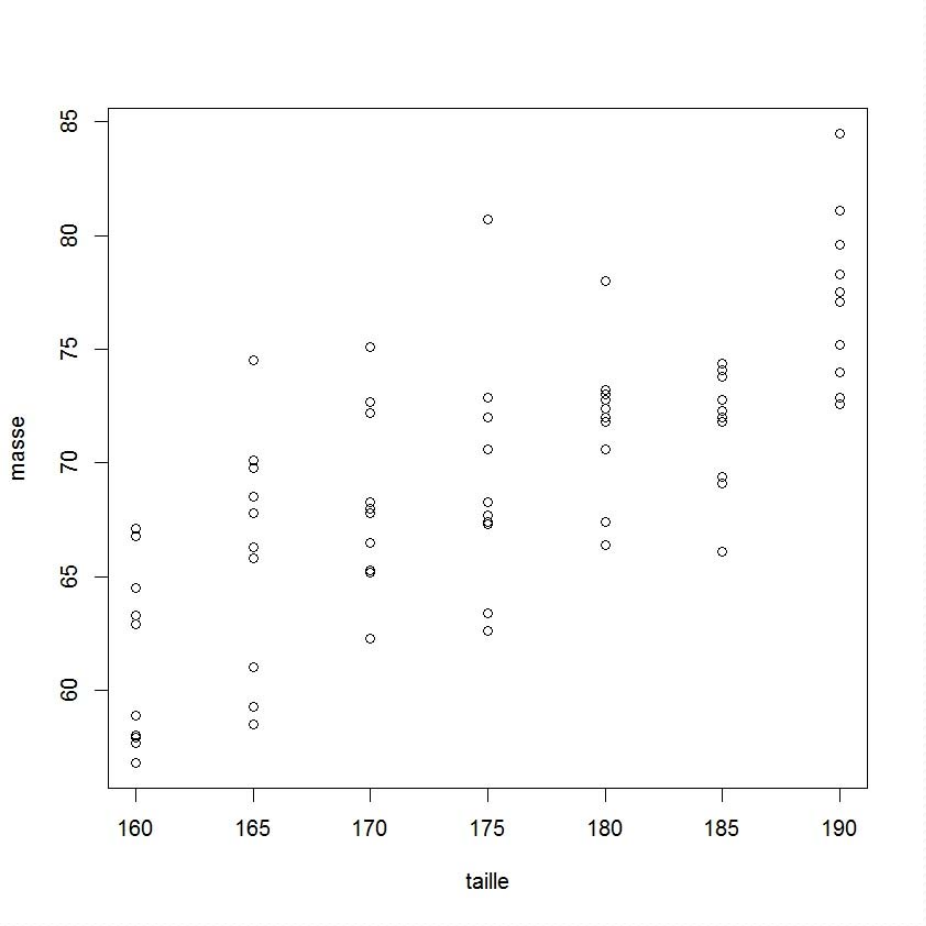
Tableau 1.3 — Taille et poids de 70 individus

Obs. i	Taille x_i	Poids y_i	Obs. i	Taille x_i	Poids y_i	Obs. i	Taille x_i	Poids y_i
1	160	57,9	25	170	65,3	49	180	71,8
2	160	58,9	26	170	65,2	50	180	72,0
3	160	63,3	27	170	68,3	51	185	74,1
4	160	56,8	28	170	62,3	52	185	74,4
5	160	66,8	29	170	67,8	53	185	72,0
6	160	64,5	30	170	66,5	54	185	66,1
7	160	67,1	31	175	67,4	55	185	69,4
8	160	58,0	32	175	67,7	56	185	71,8
9	160	62,9	33	175	62,6	57	185	73,8
10	160	57,7	34	175	70,6	58	185	69,1
11	165	68,5	35	175	72,0	59	185	72,3
12	165	69,8	36	175	68,3	60	185	72,8
13	165	58,5	37	175	72,9	61	190	72,6
14	165	66,3	38	175	63,4	62	190	81,1
15	165	65,8	39	175	80,7	63	190	78,3
16	165	61,0	40	175	67,3	64	190	72,9
17	165	74,5	41	180	67,4	65	190	79,6
18	165	59,3	42	180	70,6	66	190	77,1
19	165	67,8	43	180	72,4	67	190	84,5
20	165	70,1	44	180	73,2	68	190	74,0
21	170	72,7	45	180	72,8	69	190	77,5
22	170	75,1	46	180	66,4	70	190	75,2
23	170	68,0	47	180	73,0			
24	170	72,2	48	180	78,0			

3. Relation stochastique

- > taille<-
  rep(seq(from=160,to=190,by=5),c(10,10,10,10,10,10,10))
- > masse<-
  c(57.9,58.9,63.3,56.8,66.8,64.5,67.1,58,62.9,57.7,68.5,69
  .8,58.5,66.3,65.8,61,74.5,59.3,67.8,70.1,72.7,75.1,68,72.2,
  65.3,65.2,68.3,62.3,67.8,66.5,67.4,67.7,62.6,70.6,72,68.3
  ,72.9,63.4,80.7,67.3,67.4,70.6,72.4,73.2,72.8,66.4,73,78,
  71.8,72,74.1,74.4,72,66.1,69.4,71.8,73.8,69.1,72.3,72.8,72.
  6,81.1,78.3,72.9,79.6,77.1,84.5,74,77.5,75.2)
- > plot(masse~taille)

3. Relation stochastique



3. Relation stochastique

Commentaires :

- Plusieurs *Y* pour une même valeur de *X*.

  ⇨ Modèle linéaire déterministe inadéquat.

- Cependant *Y* augmente quand *X* augmente.

  ⇨ Modèle linéaire stochastique envisageable.

3. Relation stochastique

Définition du modèle linéaire stochastique :

$$\mu_Y(x) = \beta_0 + \beta_1 x$$

$\mu_Y(x)$: moyenne de *Y* mesurée sur tous les individus pour lesquels *X* vaut *x*.

3. Relation stochastique

Remarques :

- Comme ε, $\mu_Y(x)$ n'est ni observable, ni calculable.
- Pour calculer $\mu_Y(x)$, il faudrait recenser **tous** les individus de la population.

3. Relation stochastique

Dans la pratique :

Nous estimons la moyenne théorique $\mu_Y(x)$ par la moyenne empirique de Y définie par :

$$\bar{y}_n(x) = \frac{1}{n}\sum_{i=1}^{n} y_i(x)$$

3. Relation stochastique

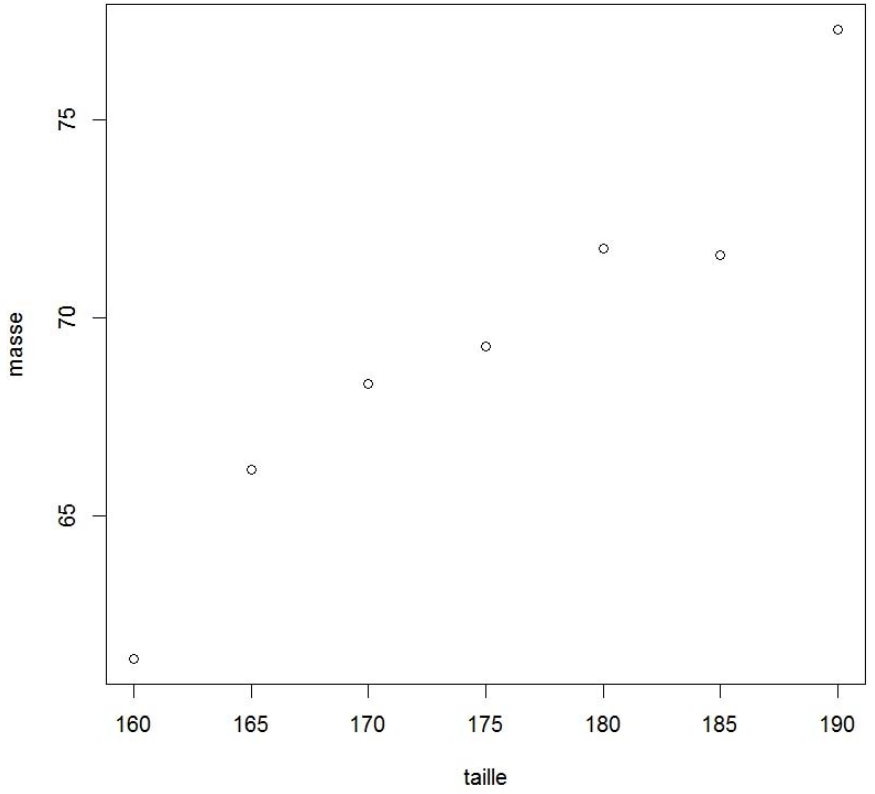
Retour à l'exemple : nous allons calculer la moyenne par tranche d'âge.

Pour cela nous allons utiliser le logiciel **R**.

```
> tableau<-data.frame(masse,taille)
> moyenne<-
   tapply(tableau$masse,tableau$taille,mean)
> moyenne
   160     165     170     175     180     185     190
61.39   66.16   68.34   69.29   71.76   71.58   77.28
```

3. Relation stochastique

Taille	Masse moyenne
160	61,39
165	66,16
170	68,34
175	69,29
180	71,76
185	71,58
190	77,28



3. Relation stochastique

La droite que nous venons de tracer s'appelle :

la droite de régression.

X et *Y* ne jouent pas un rôle identique.

X explique *Y* → *X* est une variable indépendante (ou explicative) et *Y* est une variable dépendante (ou expliquée).

3. Relation stochastique

En analyse de régression linéaire :

x_i est fixé

y_i est aléatoire

la composante aléatoire d'un y_i est le ε_i correspondant.

3. Relation stochastique

Pour l'instant, la droite de régression est inconnue.

Tout le problème est d'estimer β_0 et β_1 à partir d'un échantillon de données.

3. Relation stochastique

Choix des paramètres : droite qui approche le mieux les données

⇨ introduction de $\hat{\beta}_0$ et $\hat{\beta}_1$ qui sont des estimateurs de β_0 et de β_1.

L'estimation de la droite de régression :

$$\hat{y}(x) = \hat{\beta}_0 + \hat{\beta}_1 x$$

3. Relation stochastique

Remarques :

- $\hat{y}(x)$ est un estimateur de $\mu_Y(x)$
- Si le modèle est bon, $\hat{y}(x)$ est plus précis que

$$\bar{y}_n(x) = \frac{1}{n}\sum_{i=1}^{n} y_i(x)$$

3. Relation stochastique

Lorsque $x = x_i$, alors $\hat{y}(x_i) = \hat{y}_i$, c'est-à-dire :

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

$\hat{y}_i$ est appelée la valeur estimée par le modèle.

3. Relation stochastique

Ces valeurs estiment les quantités inobservables :

$$\varepsilon_i = y_i - \beta_0 - \beta_1 x_i$$

par les quantités observables :

$$e_i = y_i - \hat{y}_i$$

3. Relation stochastique

- Ces quantités e_i = les résidus du modèle.
- La plupart des méthodes d'estimation : estimer la droite de régression par une droite qui minimise une fonction de résidus.
- La plus connue : la méthode des moindres carrés ordinaires.

4. Méthode des moindres carrés ordinaires

Méthode : Définir des estimateurs qui minimisent la somme des carrés des résidus

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

$$= \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

4. Méthode des moindres carrés ordinaires

Les estimateurs sont donc les coordonnées du minimum de la fonction à deux variables :

$$z = f(\beta_0, \beta_1) = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

Cette fonction est appelée la **fonction objectif**.

4. Méthode des moindres carrés ordinaires

Les estimateurs correspondent aux valeurs annulant les dérivées partielles de cette fonction :

$$\frac{\partial z}{\partial \beta_0} = -2\sum (y_i - \beta_0 - \beta_1 x_i)$$

$$\frac{\partial z}{\partial \beta_1} = -2\sum x_i (y_i - \beta_0 - \beta_1 x_i)$$

4. Méthode des moindres carrés ordinaires

Les estimateurs sont les solutions du système :

$$-2\sum(y_i - \beta_0 - \beta_1 x_i) = 0$$
$$-2\sum x_i(y_i - \beta_0 - \beta_1 x_i) = 0$$

Soient :

$$(4.1) \quad \sum y_i = n\hat{\beta}_0 + \hat{\beta}_1 \sum x_i$$

$$(4.2) \quad \sum x_i y_i = \hat{\beta}_0 \sum x_i + \hat{\beta}_1 \sum x_i^2$$

4. Méthode des moindres carrés ordinaires

Nous notons :

$$\bar{x}_n = \frac{\sum x_i}{n} \text{ et } \bar{y}_n = \frac{\sum y_i}{n}$$

D'après (4.1), nous avons :

$$\hat{\beta}_0 = \bar{y}_n - \hat{\beta}_1 \bar{x}_n$$

4. Méthode des moindres carrés ordinaires

A partir de (4.2), nous avons :

$$\hat{\beta}_1 \sum x_i^2 = \sum x_i y_i - \hat{\beta}_0 n \bar{x}_n$$
$$= \sum x_i y_i - n \bar{x}_n \bar{y}_n + \hat{\beta}_1 n (\bar{x}_n)^2$$

Ainsi nous obtenons :

$$\hat{\beta}_1 = \frac{\sum x_i y_i - n \bar{x}_n \bar{y}_n}{\sum x_i^2 - n (\bar{x}_n)^2}$$

4. Méthode des moindres carrés ordinaires

Comme nous avons :

$$\sum (x_i - \bar{x}_n)(y_i - \bar{y}_n) = \sum x_i y_i - n\bar{x}_n \bar{y}_n$$

$$\sum (x_i - \bar{x}_n)^2 = \sum x_i^2 - n(\bar{x}_n)^2$$

Ainsi nous obtenons :

$$\hat{\beta}_1 = \frac{\sum (x_i - \bar{x}_n)(y_i - \bar{y}_n)}{\sum (x_i - \bar{x}_n)^2}$$

4. Méthode des moindres carrés ordinaires

Dans la pratique, nous calculons $\hat{\beta}_1$ puis $\hat{\beta}_0$

Nous obtenons une estimation de la droite de régression, appelée la **droite des moindres carrés ordinaires :**

$$\hat{y}(x) = \hat{\beta}_0 + \hat{\beta}_1 x$$

4. Méthode des moindres carrés ordinaires

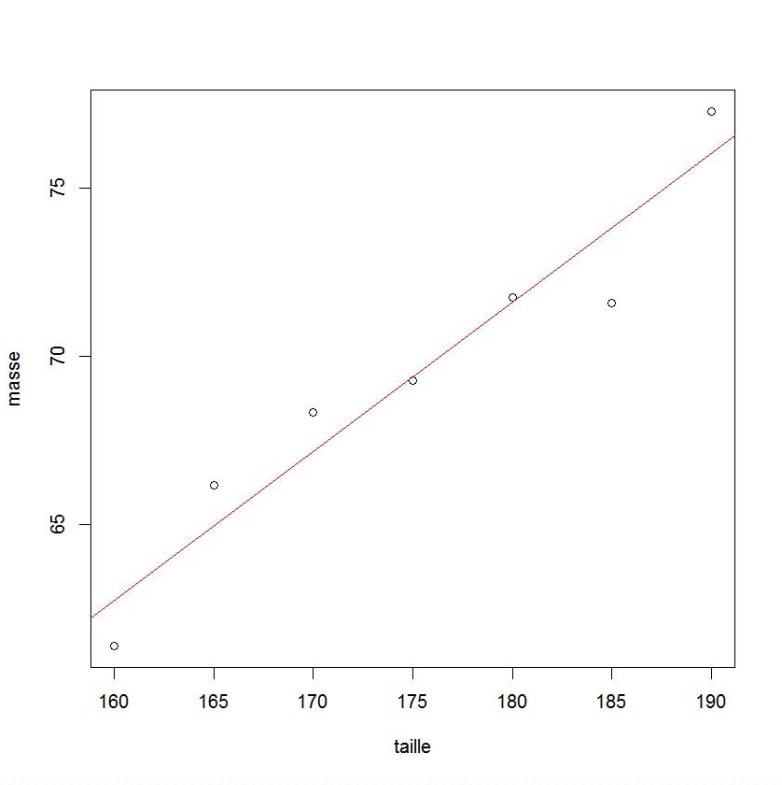
Pour obtenir les coefficients de la droite des moindres carrés et le graphique qui superpose à la fois les points et la droite, voici les lignes de commande qu'il faut effectuer :

```
> modele1<-lm(masse~taille)
> abline(coef(modele1),col="red")
```

- coef(modele1)
- (Intercept) taille

  -8.0125000 0.4423571

4. Méthode des moindres carrés ordinaires



5. Variation expliquée et inexpliquée

But d'un modèle de régression linéaire :

expliquer une partie de la variation de la variable expliquée *Y*.

La variation de *Y* vient du fait de sa dépendance à la variable explicative *X*.

➡ **Variation expliquée par le modèle.**

5. Variation expliquée et inexpliquée

Dans **l'exemple « taille-masse »**, nous avons remarqué que lorsque nous mesurons *Y* avec une même valeur de *X*, nous observons une certaine variation sur *Y*.

➡ **Variation inexpliquée par le modèle.**

5. Variation expliquée et inexpliquée

Variation totale de *Y*

= Variation expliquée par le modèle

+ Variation inexpliquée par le modèle

5. Variation expliquée et inexpliquée

Pour mesurer la variation de *Y* : nous introduisons $\bar{y}_n$

$$(y_i - \bar{y}_n) = (\hat{y}_i - \bar{y}_n) + (y_i - \hat{y}_i)$$

Différence expliquée par le modèle

Différence inexpliquée par le modèle ou résidu du modèle

5. Variation expliquée et inexpliquée

Pourquoi la méthode des moindres carrés ?

- Un propriété remarquable : elle conserve une telle décomposition en considérant la somme des carrés de ces différences :

$$\sum (y_i - \bar{y}_n)^2 = \sum (\hat{y}_i - \bar{y}_n)^2 + \sum (y_i - \hat{y}_i)^2$$

5. Variation expliquée et inexpliquée

$$\sum (y_i - \bar{y}_n)^2 = \sum (\hat{y}_i - \bar{y}_n)^2 + \sum (y_i - \hat{y}_i)^2$$

Somme des carrés totale

(SC_{tot})

Somme des carrés due à la régression

(SC_{reg})

Somme des carrés des résidus

(SC_{res})

5. Variation expliquée et inexpliquée

Mesure du pourcentage de la variation totale expliquée par le modèle :

Introduction d'un **coefficient de détermination**

$$R^2 = \frac{\text{Variation expliquée}}{\text{Variation totale}} = \frac{SC_{reg}}{SC_{tot}}$$

5. Variation expliquée et inexpliquée

Quelques remarques :

- R^2 est compris entre 0 et 1.
- R^2 =1 : cas où les données sont parfaitement alignées (comme c'est le cas pour un modèle déterministe).
- R^2 =0 : cas où la variation de Y n'est pas due à la variation de X. Les données ne sont pas du tout alignées.
- Plus R^2 est proche de 1, plus les données sont alignées sur la droite de régression.