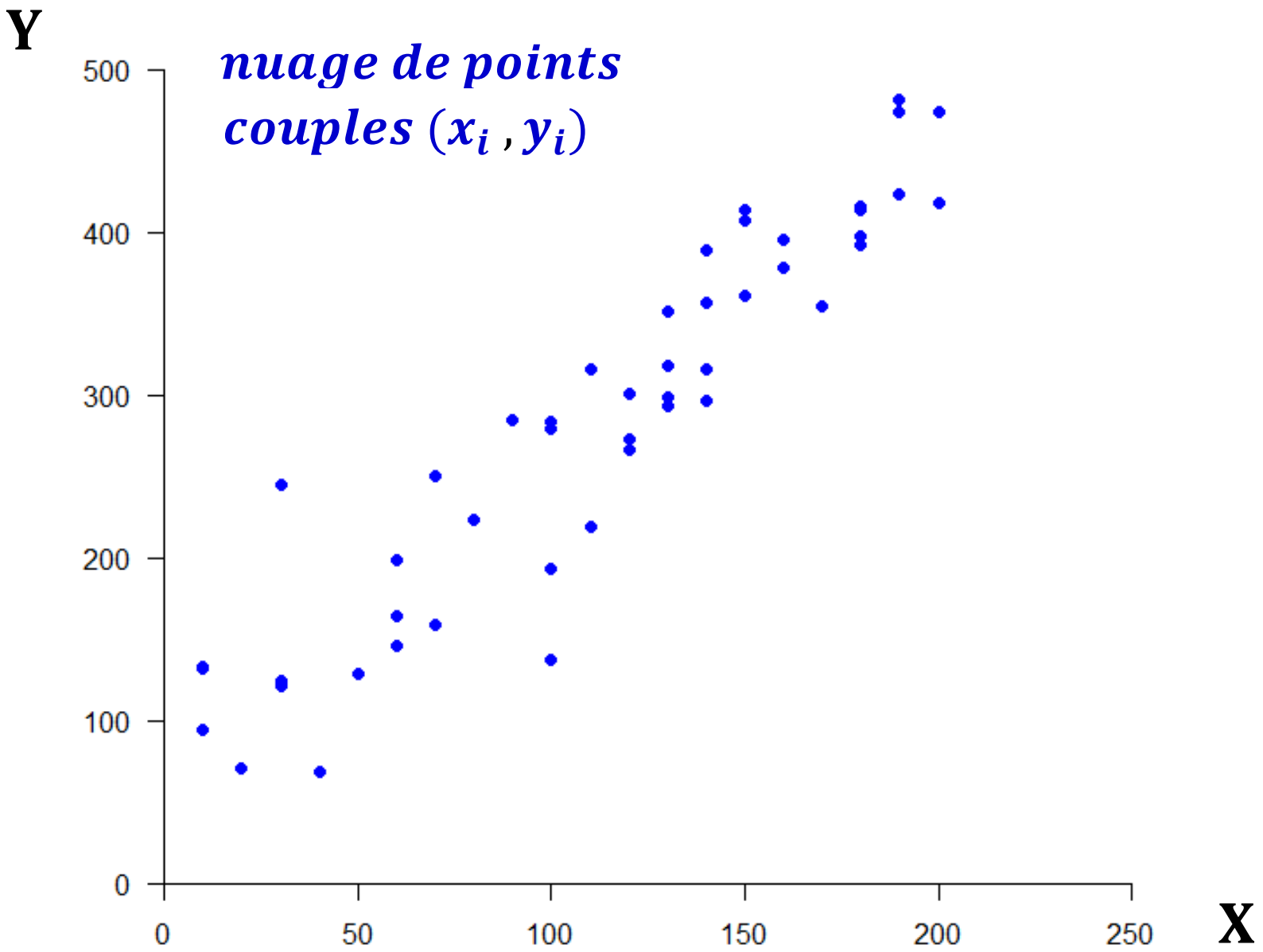
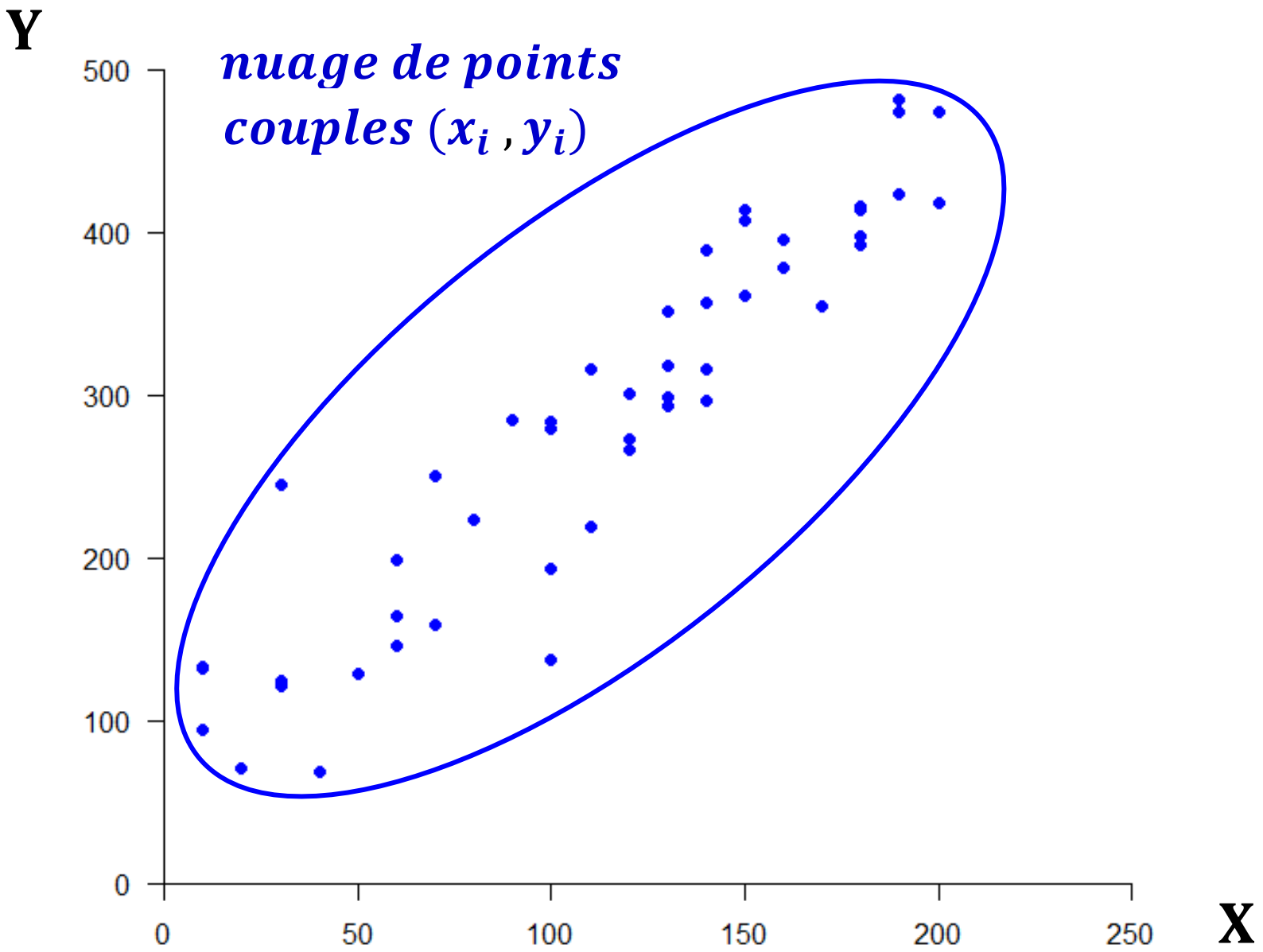
La Régression Linéaire : une brève présentation et pratique avec Rcmdr

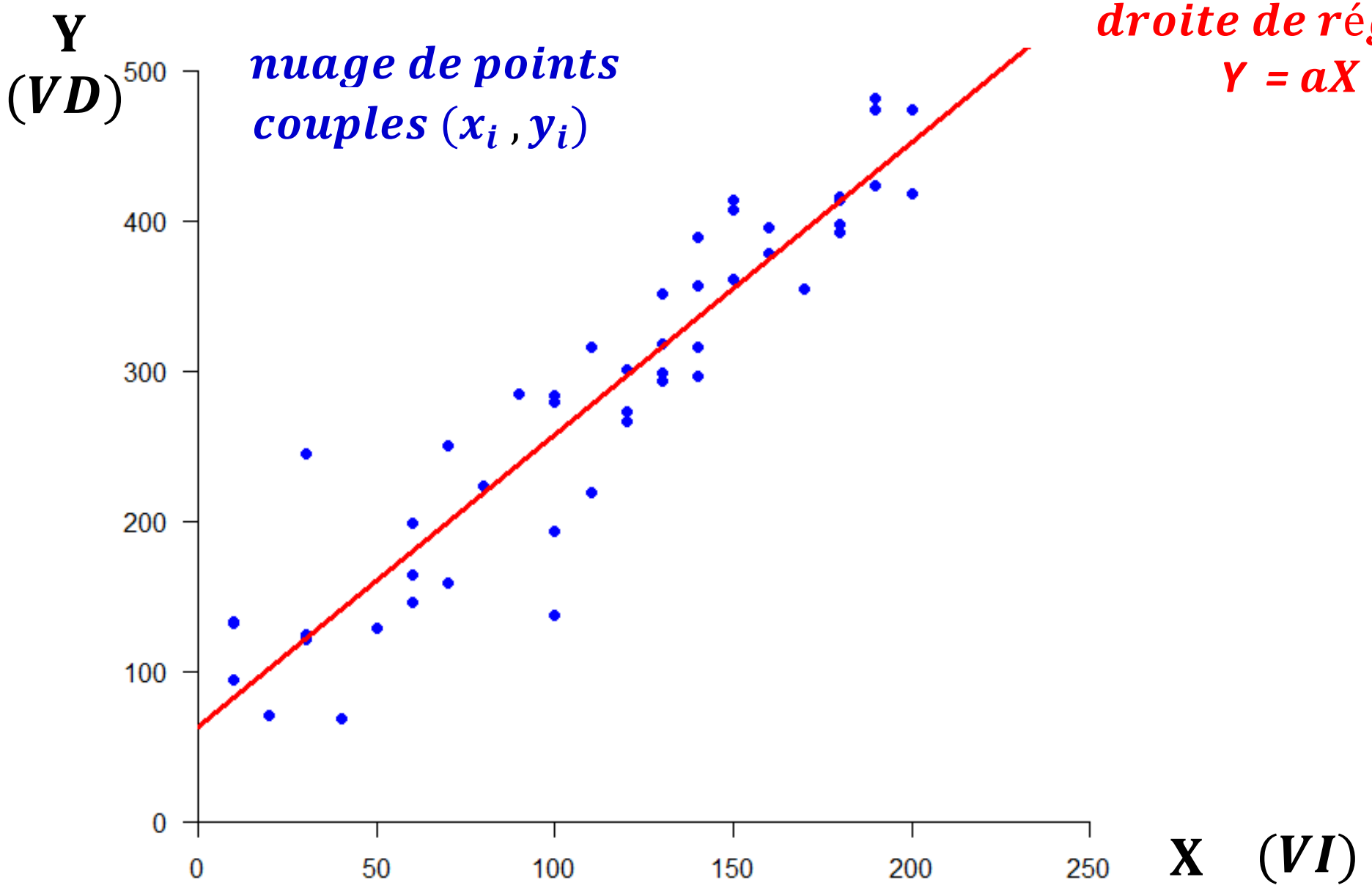
La Régression Linéaire : quelques repères



La Régression Linéaire : quelques repères



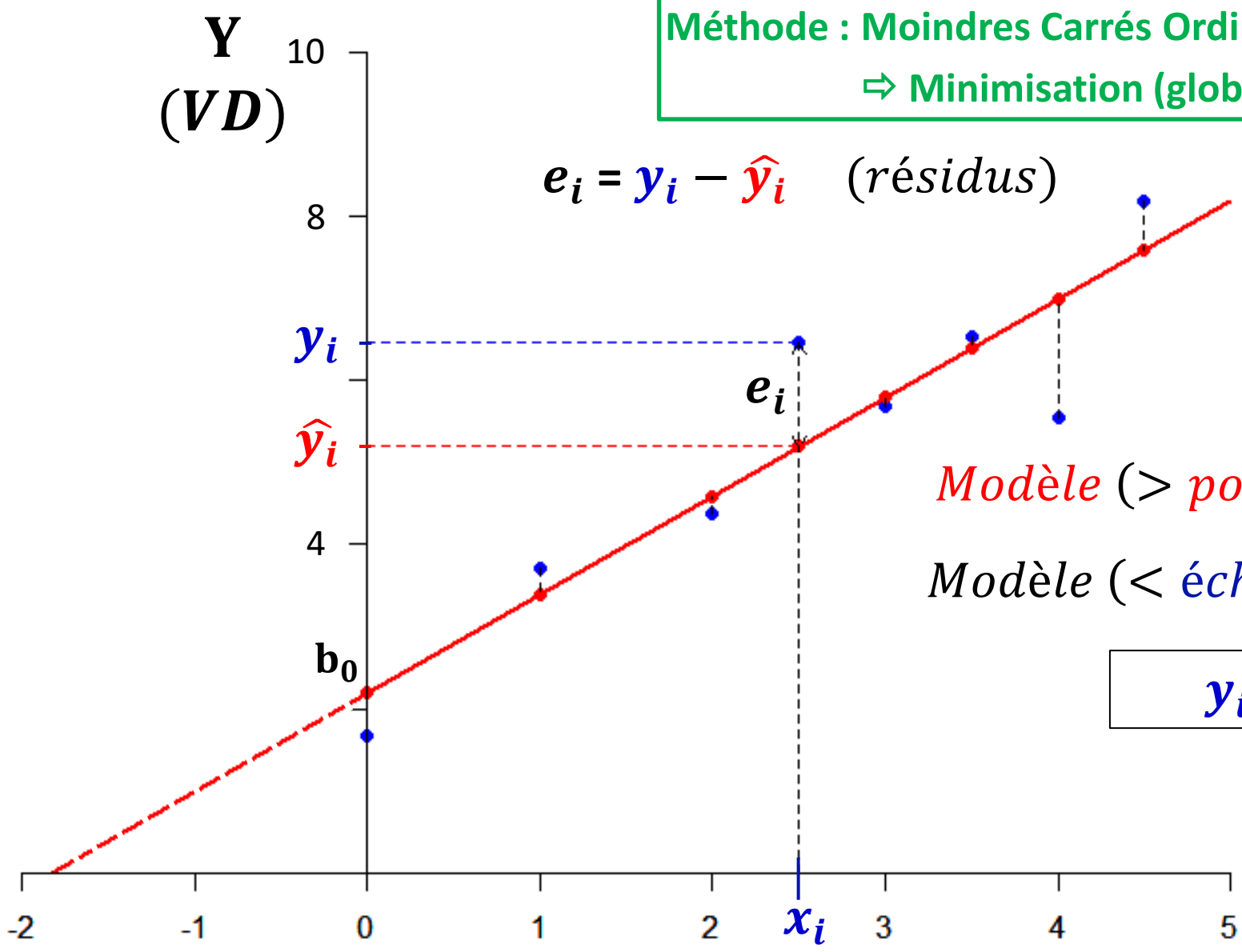
La Régression Linéaire : quelques repères



La Régression Linéaire : quelques repères

Méthode : Moindres Carrés Ordinaires MCO / OLS (> calcul matriciel)

⇨ Minimisation (globale) des erreurs/résidus



$e_i = y_i - \hat{y}_i$ (résidus)

droite de régression

$Y = b_0 + b_1 X$

$\hat{y}_i = b_0 + b_1 x_i$

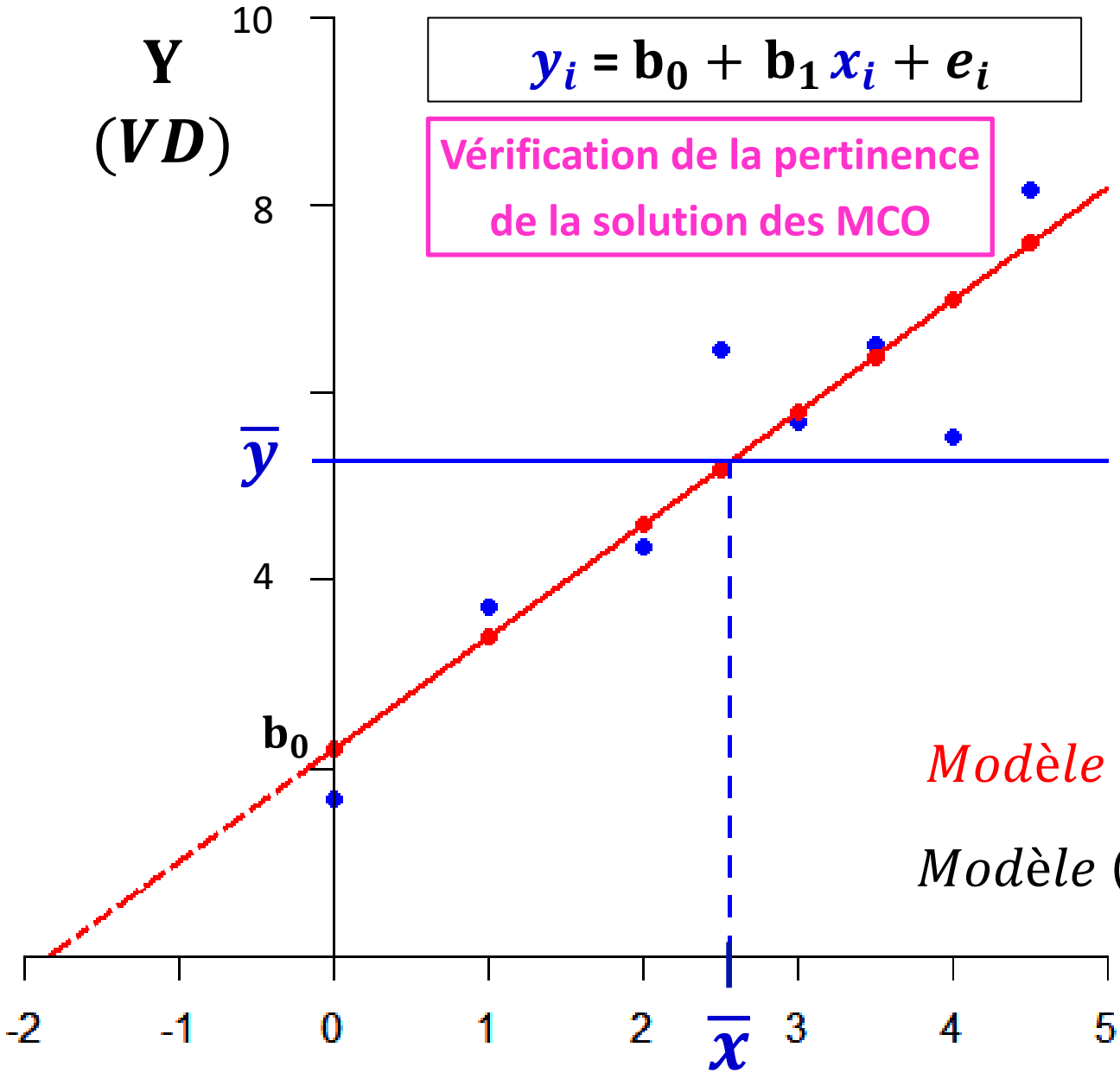
Modèle (> population) : $Y = \beta_0 + \beta_1 X + \varepsilon$

Modèle (< échantillon) : $Y = b_0 + b_1 X + E$

$$y_i = b_0 + b_1 x_i + e_i$$

$b_0 = \widehat{\beta_0}$; $b_1 = \widehat{\beta_1}$

La Régression Linéaire : quelques repères



$$y_i = b_0 + b_1 x_i + e_i$$

Vérification de la pertinence de la solution des MCO

droite de régression

$$Y = b_0 + b_1 X$$

$$\hat{y}_i = b_0 + b_1 x_i$$

Tests sur les coefficients b :

(tests de Student)

Ho : $\beta_1 = 0$; **H1 :** $\beta_1 \neq 0$

Ho : $\beta_0 = 0$; **H1 :** $\beta_0 \neq 0$

Modèle (> population) : $Y = \beta_0 + \beta_1 X + \varepsilon$

Modèle (< échantillon) : $Y = b_0 + b_1 X + E$

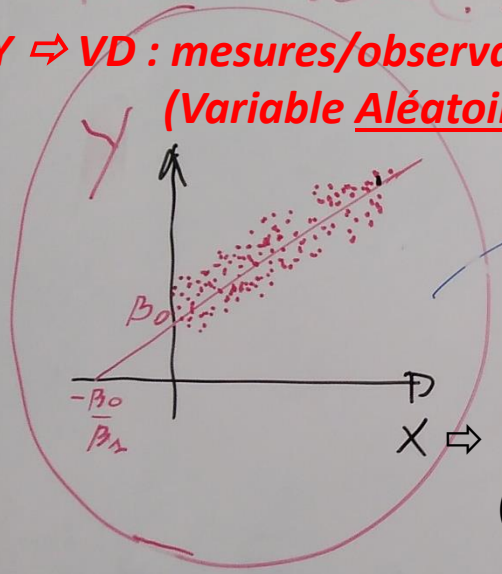
$b_0 = \widehat{\beta_0}$; $b_1 = \widehat{\beta_1}$

Régression linéaire : quelques repères

Y : VA d'étude (normale)

VRAI MODÈLE (??) ?

Y ⇨ VD : mesures/observations (Variable Aléatoire)



X ⇨ VI : variable(s) fixée(s) (explicatives)

population

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

Modèle (explicatif)

régression population (ne sert jamais déterminable précisément)

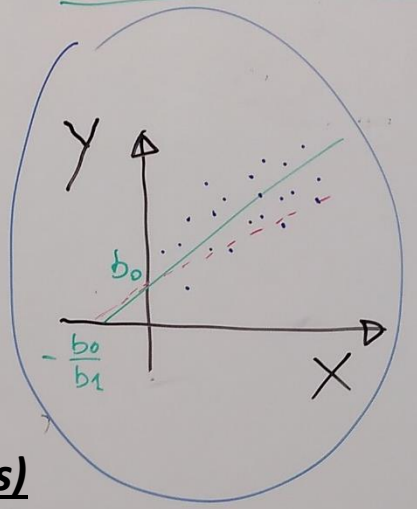
$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$

résidu population

$(x_i, y_i, \varepsilon_i)$

Modèle obtenu (approche du modèle réel inconnu)

Modèle déduit de nos observations (x_i, y_i)



échantillon (n)

$$Y = b_0 + b_1 \cdot X$$

$y_i = b_0 + b_1 x_i + e_i$

(x_i, y_i, e_i)

β_1 ?

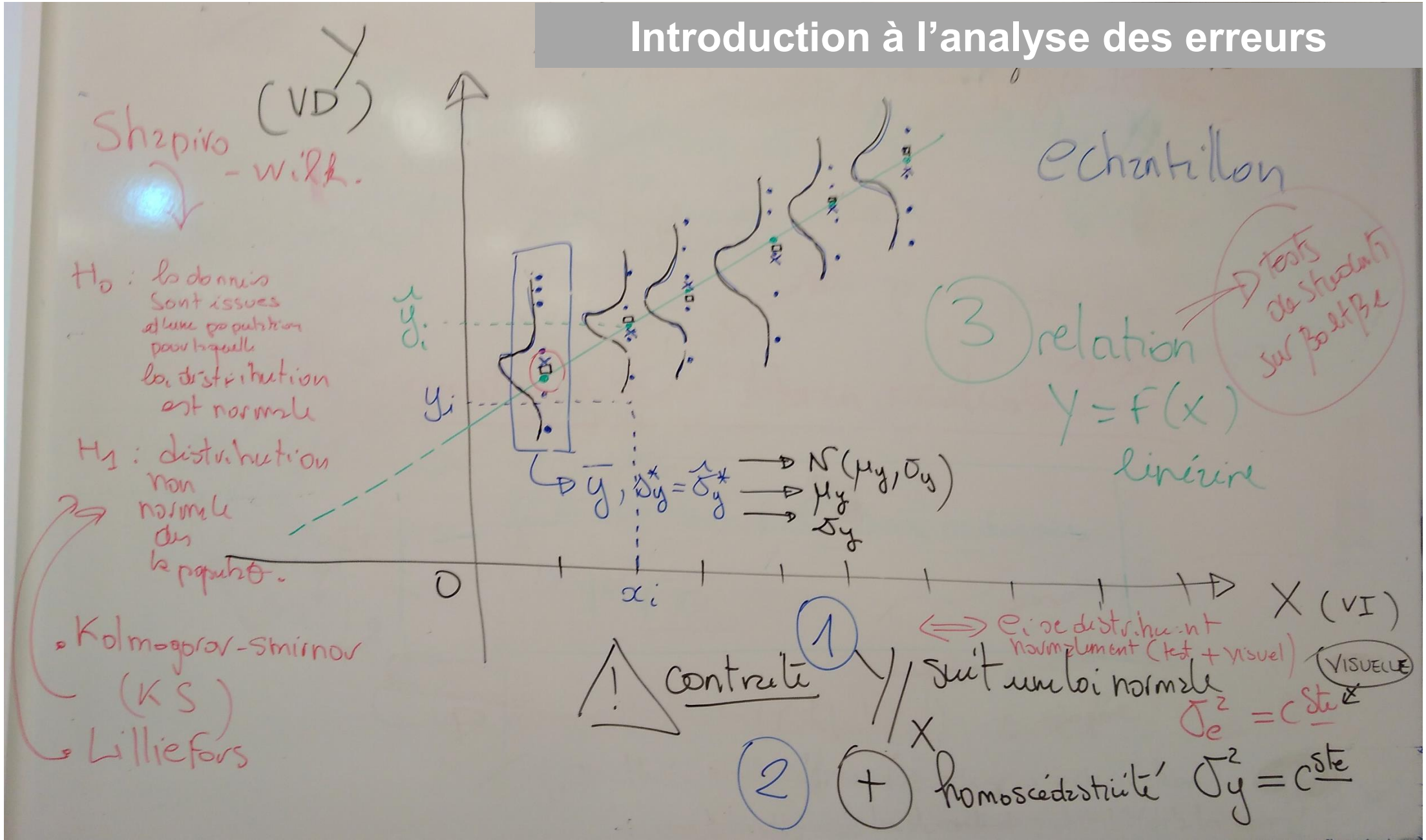
$\hookrightarrow b_1 \pm t_{(n-2)} \cdot \frac{\hat{\sigma}^*_{\beta_1}}{\sqrt{n}}$

$\begin{cases} b_0 = \hat{\beta}_0 \\ b_1 = \hat{\beta}_1 \end{cases}$

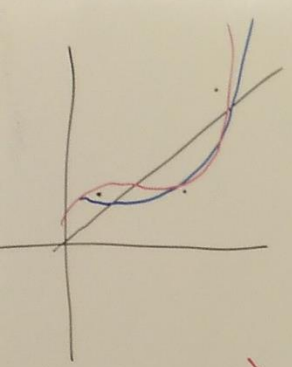
estimations ponctuelles des vrais coefficients (inconnus)

La Régression Linéaire : quelques repères

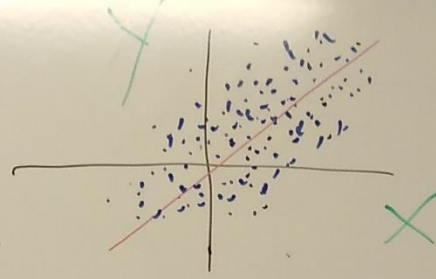
Introduction à l'analyse des erreurs



Régression linéaire : suite de l'analyse



Nuage de point



$\hat{\rho} = r \simeq 0,3$

Pearson $R^2 \simeq 0,09$

MCO : méthode des Moindres Carrés Ordinaires (Maths)

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

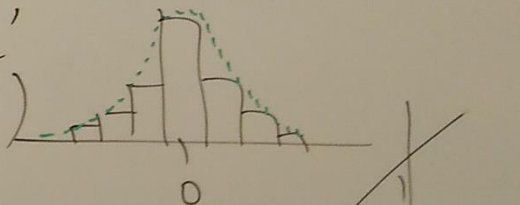
X explique-t-il Y ou bien Y expliqué par μ_Y

droite réaliste ??

① droite de régression $Y = b_0 + b_1 X$

② analyse des résidus e_i $\quad y_i = b_0 + b_1 x_i + e_i$

+ test normalité (Shapiro-Wilk)



③ B_1 significativement différent de 0 ?

test de Student sur le coefficient

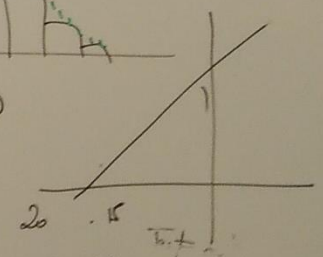
$t_0 = \dfrac{b_1 - 0}{\sqrt{\dfrac{\sigma^{*2}_{b_1}}{n}}}$ $\quad \downarrow H_0$

④ R^2 part de la variance expliquée par le modèle

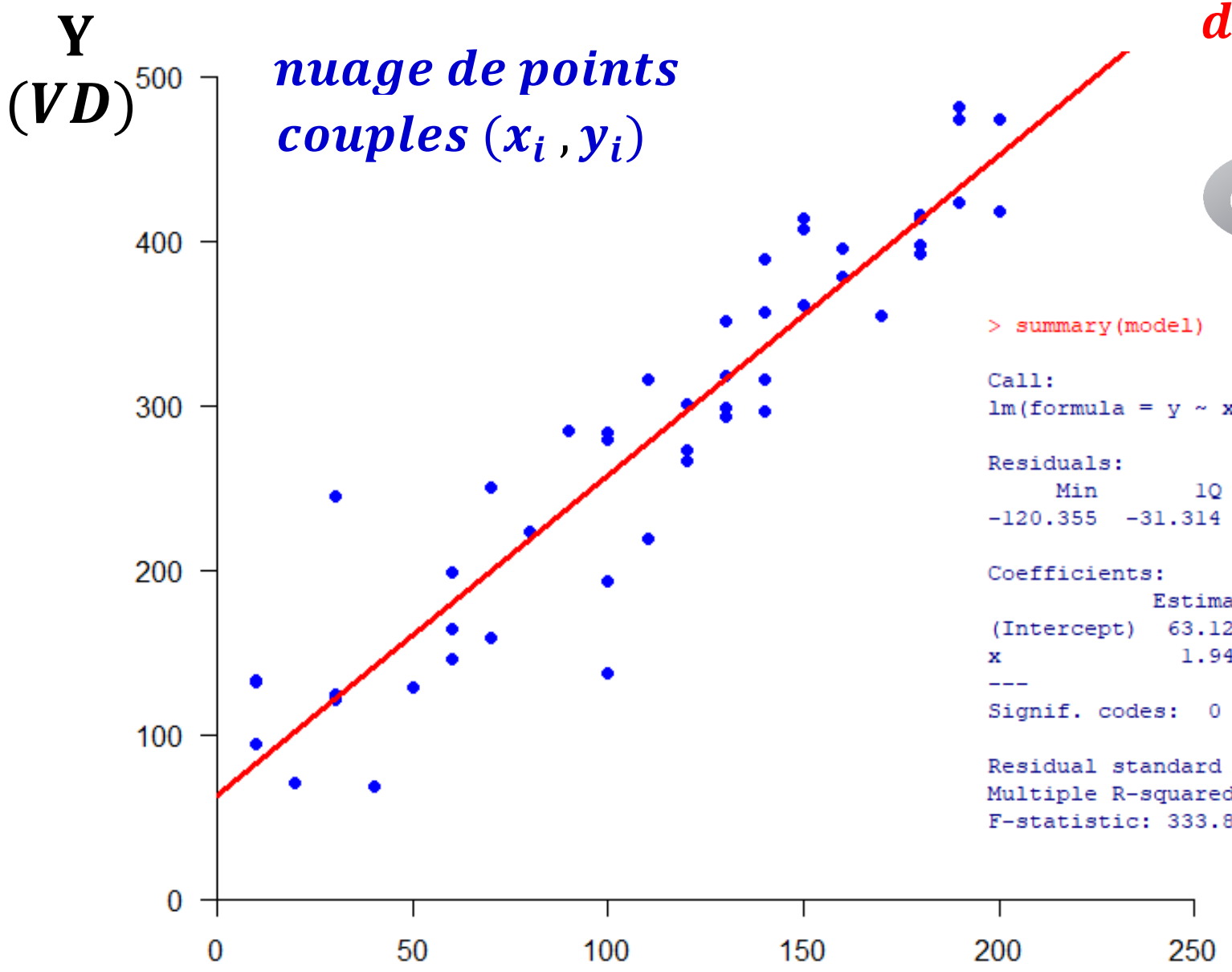
$H_0 : B_1 = 0$

$H_1 : B_1 \neq 0$

$\Longleftrightarrow$ test F sur la variance (Fisher)



La Régression Linéaire avec un logiciel

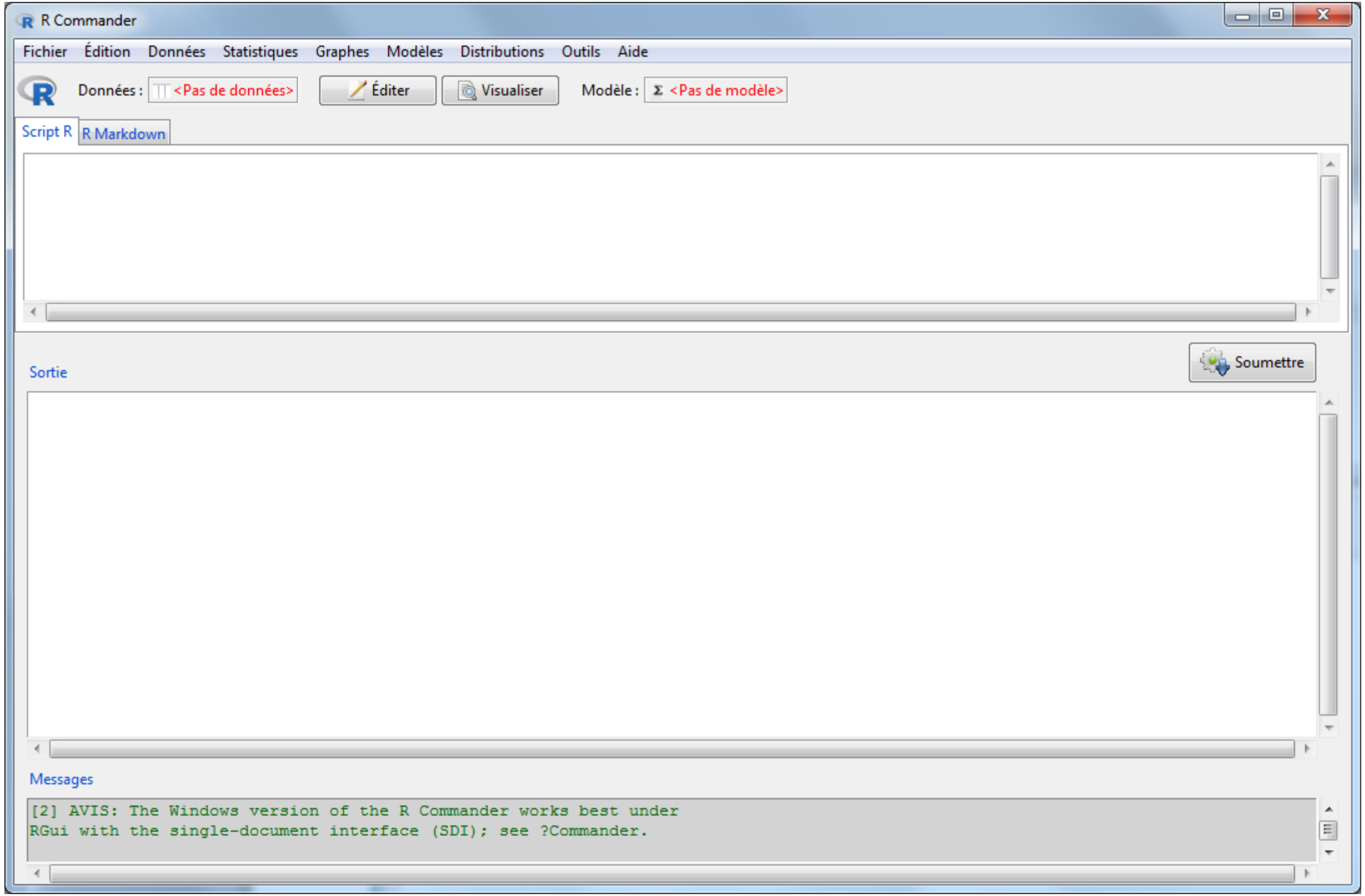


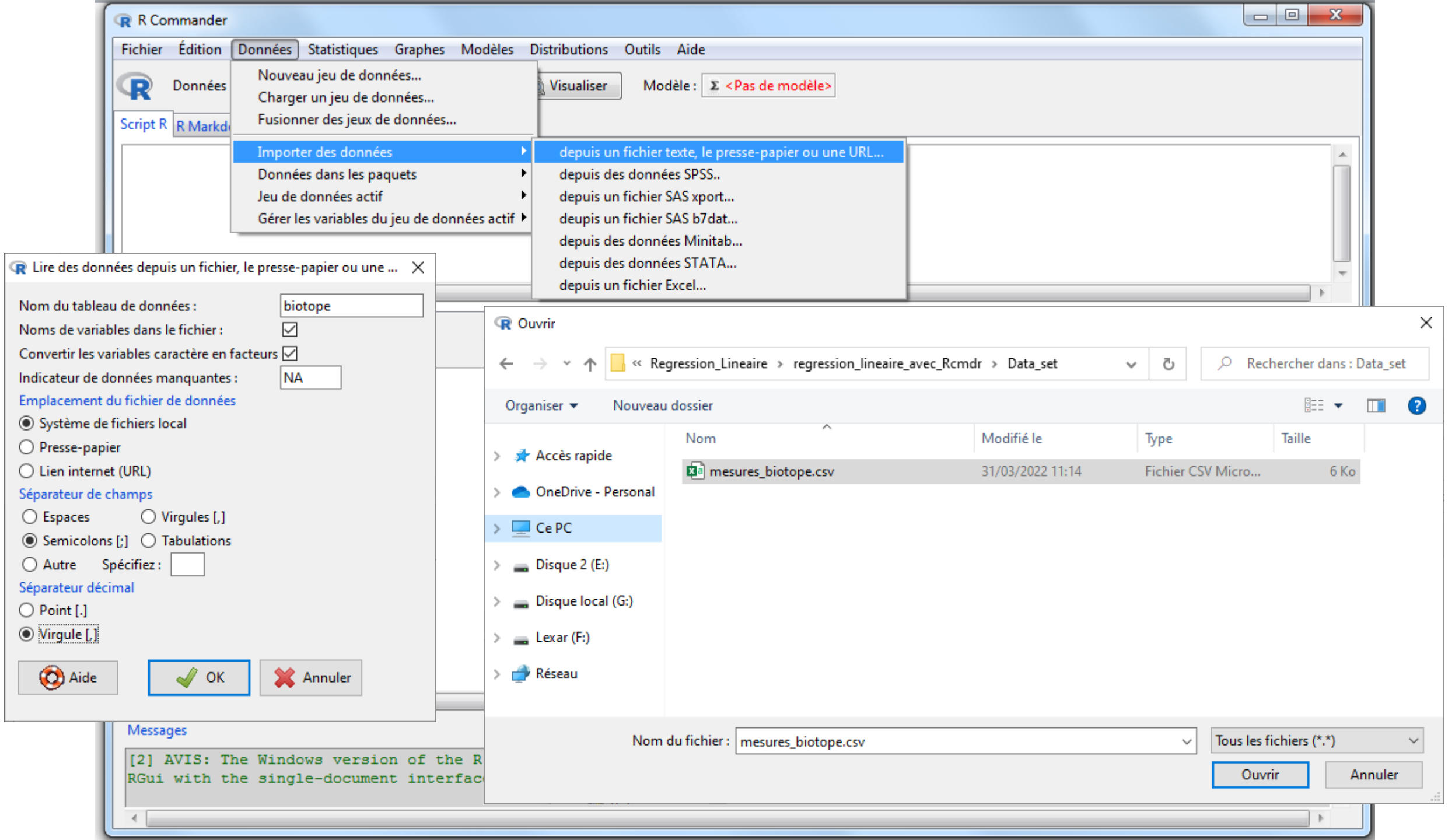
nuage de points
couples (x_i, y_i)

droite de régression
$Y = aX + b$

```
> summary(model)

Call:
lm(formula = y ~ x)

Residuals:
     Min       1Q   Median       3Q      Max
-120.355  -31.314    2.038   24.539  123.736

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  63.1205    13.1979   4.783 1.68e-05 ***
x             1.9460     0.1065  18.269  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 42.37 on 48 degrees of freedom
Multiple R-squared:  0.8743,    Adjusted R-squared:  0.8716
F-statistic: 333.8 on 1 and 48 DF,  p-value: < 2.2e-16
```

La Régression Linéaire : pratique avec Rcmdr


R Commander
Fichier Édition Données Statistiques Graphes Modèles Distributions Outils Aide
Données : <Pas de données>
Éditer
Visualiser
Modèle : Σ <Pas de modèle>
Script R
R Markdown
Soumettre
Sortie
Messages
[2] AVIS: The Windows version of the R Commander works best under
RGui with the single-document interface (SDI); see ?Commander.


R Commander
Fichier Édition Données Statistiques Graphes Modèles Distributions Outils Aide
Données
Visualiser
Modèle : <Pas de modèle>
Script R
Nouveau jeu de données...
Charger un jeu de données...
Fusionner des jeux de données...
Importer des données
Données dans les paquets
Jeu de données actif
Gérer les variables du jeu de données actif
depuis un fichier texte, le presse-papier ou une URL...
depuis des données SPSS..
depuis un fichier SAS xport...
deupis un fichier SAS b7dat...
depuis des données Minitab...
depuis des données STATA...
depuis un fichier Excel...
Lire des données depuis un fichier, le presse-papier ou une ...
Nom du tableau de données : biotope
Noms de variables dans le fichier :
Convertir les variables caractère en facteurs
Indicateur de données manquantes : NA
Emplacement du fichier de données
Système de fichiers local
Presse-papier
Lien internet (URL)
Séparateur de champs
Espaces Virgules [,]
Semicolons [;] Tabulations
Autre Spécifiez :
Séparateur décimal
Point [.]
Virgule [,]
Aide OK Annuler
Ouvrir
« Regression_Lineaire › regression_lineaire_avec_Rcmdr › Data_set
Rechercher dans : Data_set
Organiser Nouveau dossier
Accès rapide
OneDrive - Personal
Ce PC
Disque 2 (E:)
Disque local (G:)
Lexar (F:)
Réseau
Nom Modifié le Type Taille
mesures_biotope.csv 31/03/2022 11:14 Fichier CSV Micro... 6 Ko
Nom du fichier : mesures_biotope.csv
Tous les fichiers (*.*)
Ouvrir Annuler
Messages
[2] AVIS: The Windows version of the R
RGui with the single-document interfac

Le fichier : mesures_biotope.csv

	A	B	C	D	E	F
1	Y	X1	X2	X3	Equipe	
2	41,1	230,1	37,8	69,2	Equipe1	
3	29,4	44,5	39,3	45,1	Equipe3	
4	28,3	17,2	45,9	69,3	Equipe1	
5	37,5	151,5	41,3	58,5	Equipe3	
6	31,9	180,8	10,8	58,4	Equipe2	
7	26,2	8,7	48,9	75	Equipe2	
8	30,8	57,5	32,8	23,5	Equipe2	
9	32,2	120,2	19,6	11,6	Equipe1	
10	23,8	8,6	2,1	1	Equipe2	
11	29,6	199,8	2,6	21,2	Equipe3	
12	27,6	66,1	5,8	24,2	Equipe1	
13	36,4	214,7	24	4	Equipe3	
14	28,2	23,8	35,1	65,9	Equipe2	
15	28,7	97,5	7,6	7,2	Equipe1	
16	38	204,1	32,9	46	Equipe2	
17	41,4	195,4	47,7	52,9	Equipe2	
18	31,5	67,8	36,6	114	Equipe1	
19	43,4	281,4	39,6	55,8	Equipe3	
20	30,3	69,2	20,5	18,3	Equipe3	
21	33,6	147,3	23,9	19,1	Equipe3	
22	37	218,4	27,7	53,4	Equipe3	
23	31,5	237,4	5,1	23,5	Equipe3	
24	24,6	13,2	15,9	49,6	Equipe3	
25	34,5	228,3	16,9	26,2	Equipe2	
26	28,7	62,3	12,6	18,3	Equipe2	

mesures_biotope

mesures_biotope.csv - Bloc-notes

Fichier Edition Format Affichage Aide

```
Y;X1;X2;X3;Equipe
41,1;230,1;37,8;69,2;Equipe1
29,4;44,5;39,3;45,1;Equipe3
28,3;17,2;45,9;69,3;Equipe1
37,5;151,5;41,3;58,5;Equipe3
31,9;180,8;10,8;58,4;Equipe2
26,2;8,7;48,9;75;Equipe2
30,8;57,5;32,8;23,5;Equipe2
32,2;120,2;19,6;11,6;Equipe1
23,8;8,6;2,1;1;Equipe2
29,6;199,8;2,6;21,2;Equipe3
27,6;66,1;5,8;24,2;Equipe1
36,4;214,7;24;4;Equipe3
28,2;23,8;35,1;65,9;Equipe2
28,7;97,5;7,6;7,2;Equipe1
38;204,1;32,9;46;Equipe2
41,4;195,4;47,7;52,9;Equipe2
31,5;67,8;36,6;114;Equipe1
43,4;281,4;39,6;55,8;Equipe3
30,3;69,2;20,5;18,3;Equipe3
33,6;147,3;23,9;19,1;Equipe3
37;218,4;27,7;53,4;Equipe3
31,5;237,4;5,1;23,5;Equipe3
24,6;13,2;15,9;49,6;Equipe3
34,5;228,3;16,9;26,2;Equipe2
28,7;62,3;12,6;18,3;Equipe2
31;262,9;3,5;19,5;Equipe1
```

Ln 1, Col 1 150% Windows (CRLF) UTF-8

R Commander

Fichier Édition Données Statistiques Graphes Modèles Distributions Outils Aide

Données : biotope | Éditer | Visualiser | Modèle : <Pas de modèle>

Script R | R Markdown

```
biotope <-
  read.table("F:/UE-Bio-Data_sciences/Machine_learning/Regression_Lineaire/regression_lineaire_avec_Rcmdr/Data_set/mesures_biotope.csv",
   header=TRUE, stringsAsFactors=TRUE, sep=";", na.strings="NA", dec=",",
  strip.white=TRUE)
```

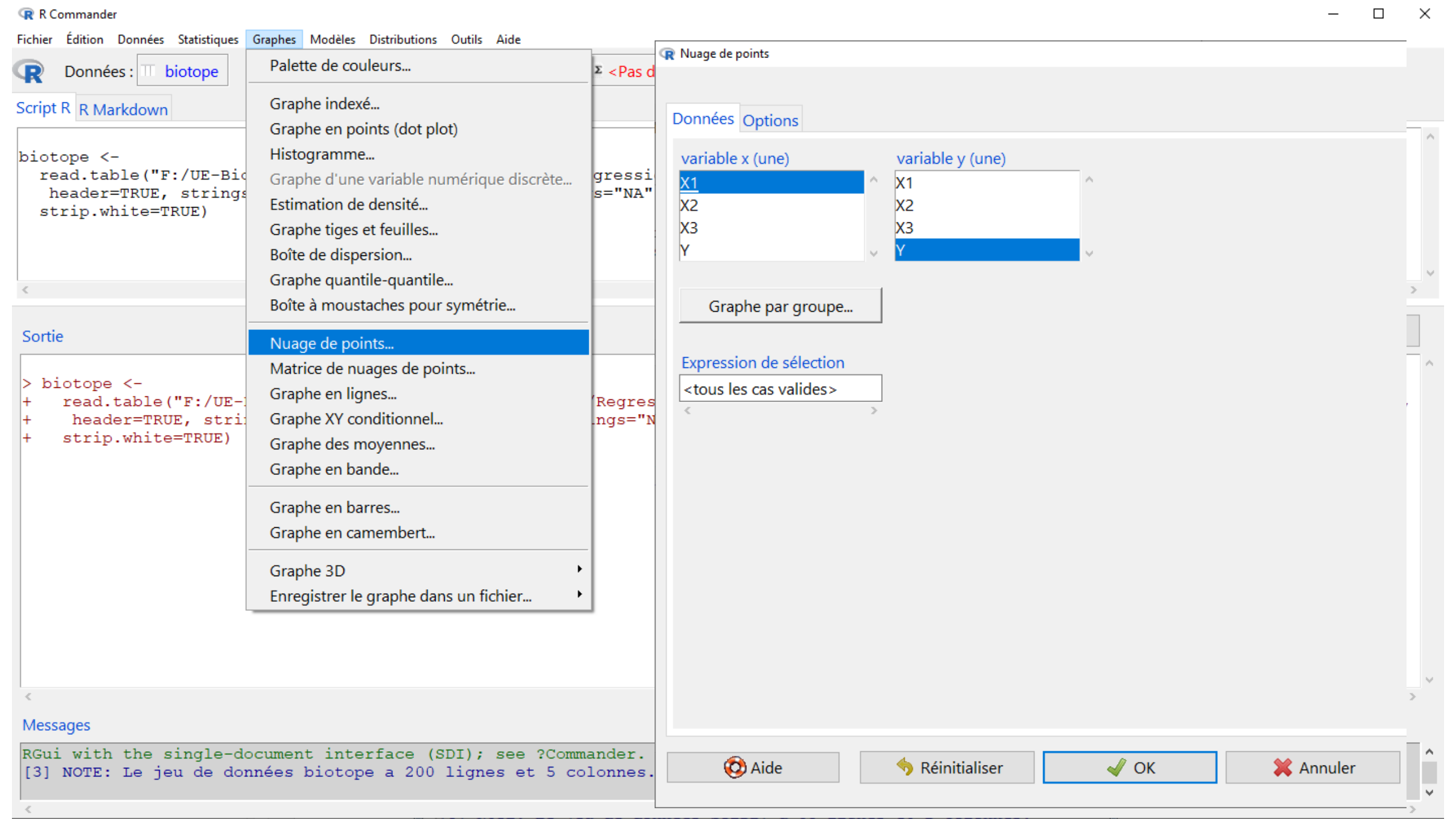
Sortie

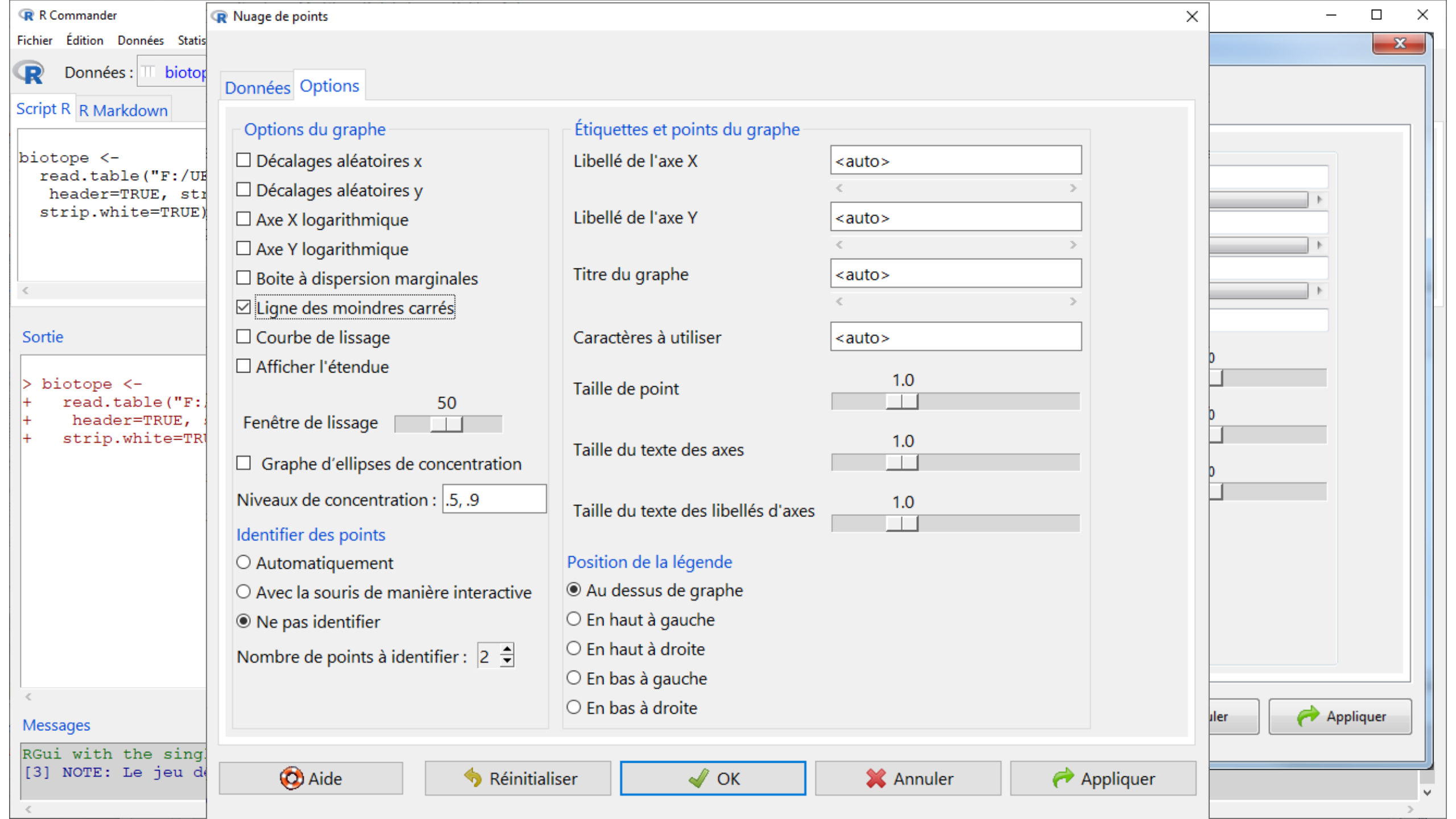
Soumettre

```
> biotope <-
+   read.table("F:/UE-Bio-Data_sciences/Machine_learning/Regression_Lineaire/regression_lineaire_avec_Rcmdr/Data_set/mesures_biotope.csv",
+    header=TRUE, stringsAsFactors=TRUE, sep=";", na.strings="NA", dec=",",
+   strip.white=TRUE)
```

Messages

```
RGui with the single-document interface (SDI); see ?Commander.
[3] NOTE: Le jeu de données biotope a 200 lignes et 5 colonnes.
```


R Commander
Fichier Édition Données Statistiques Graphes Modèles Distributions Outils Aide
Données : biotope
Script R R Markdown
biotope <-
read.table("F:/UE-Bio
header=TRUE, strings
strip.white=TRUE)
Sortie
> biotope <-
+ read.table("F:/UE-
+ header=TRUE, stri
+ strip.white=TRUE)
Messages
RGui with the single-document interface (SDI); see ?Commander.
[3] NOTE: Le jeu de données biotope a 200 lignes et 5 colonnes.
Palette de couleurs...
Graphe indexé...
Graphe en points (dot plot)
Histogramme...
Graphe d'une variable numérique discrète...
Estimation de densité...
Graphe tiges et feuilles...
Boîte de dispersion...
Graphe quantile-quantile...
Boîte à moustaches pour symétrie...
Nuage de points...
Matrice de nuages de points...
Graphe en lignes...
Graphe XY conditionnel...
Graphe des moyennes...
Graphe en bande...
Graphe en barres...
Graphe en camembert...
Graphe 3D
Enregistrer le graphe dans un fichier...
Nuage de points
Données Options
variable x (une)
X1
X2
X3
Y
variable y (une)
X1
X2
X3
Y
Graphe par groupe...
Expression de sélection
<tous les cas valides>
Aide
Réinitialiser
OK
Annuler


R Commander
Fichier Édition Données Statis
Données : biotop
Script R R Markdown
biotope <-
read.table("F:/UE
header=TRUE, str
strip.white=TRUE)
Sortie
> biotope <-
+ read.table("F:
+ header=TRUE,
+ strip.white=TR
Messages
RGui with the sing
[3] NOTE: Le jeu d
Nuage de points
Données Options
Options du graphe
Décalages aléatoires x
Décalages aléatoires y
Axe X logarithmique
Axe Y logarithmique
Boite à dispersion marginales
Ligne des moindres carrés
Courbe de lissage
Afficher l'étendue
Fenêtre de lissage
50
Graphe d'ellipses de concentration
Niveaux de concentration : .5, .9
Identifier des points
Automatiquement
Avec la souris de manière interactive
Ne pas identifier
Nombre de points à identifier : 2
Étiquettes et points du graphe
Libellé de l'axe X
<auto>
Libellé de l'axe Y
<auto>
Titre du graphe
<auto>
Caractères à utiliser
<auto>
Taille de point
1.0
Taille du texte des axes
1.0
Taille du texte des libellés d'axes
1.0
Position de la légende
Au dessus de graphe
En haut à gauche
En haut à droite
En bas à gauche
En bas à droite
Aide
Réinitialiser
OK
Annuler
Appliquer
Appliquer

Nuage de points

Données | Options

Options du graphe

- ☐ Décalages aléatoires x
- ☐ Décalages aléatoires y
- ☐ Axe X logarithmique
- ☐ Axe Y logarithmique
- ☐ Boite à dispersion marginales
- ☑ Ligne des moindres carrés
- ☐ Courbe de lissage
- ☐ Afficher l'étendue

Fenêtre de lissage : 50

- ☐ Graphe d'ellipses de concentration

Niveaux de concentration : .5, .9

Identifier des points

- ○ Automatiquement
- ○ Avec la souris de manière interactive
- ◉ Ne pas identifier

Nombre de points à identifier : 2

Étiquettes et points du graphe

Libellé de l'axe X : <auto>

Libellé de l'axe Y : <auto>

Titre du graphe : <auto>

Caractères à utiliser : <auto>

Taille de point : 1.0

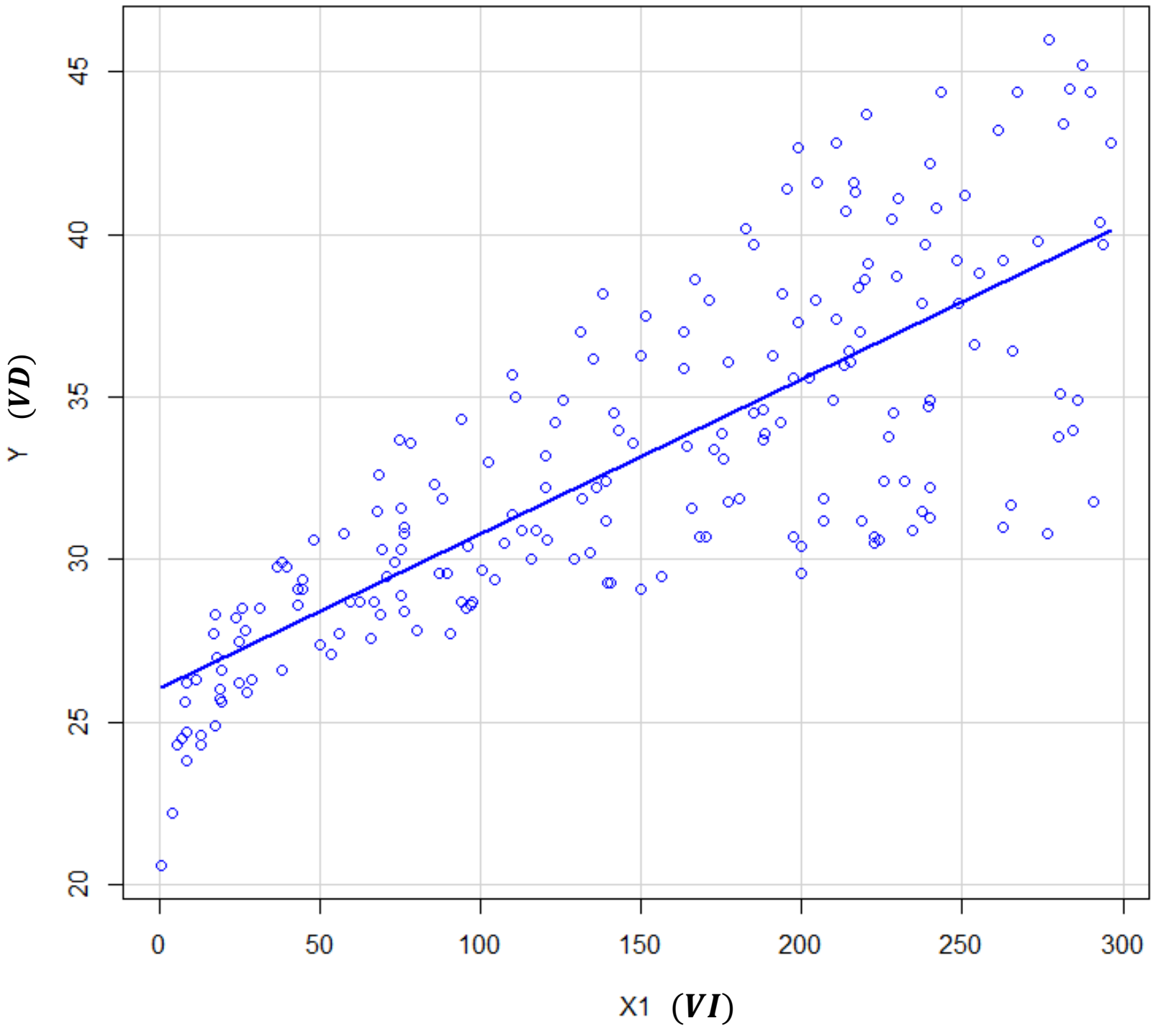
Taille du texte des axes : 1.0

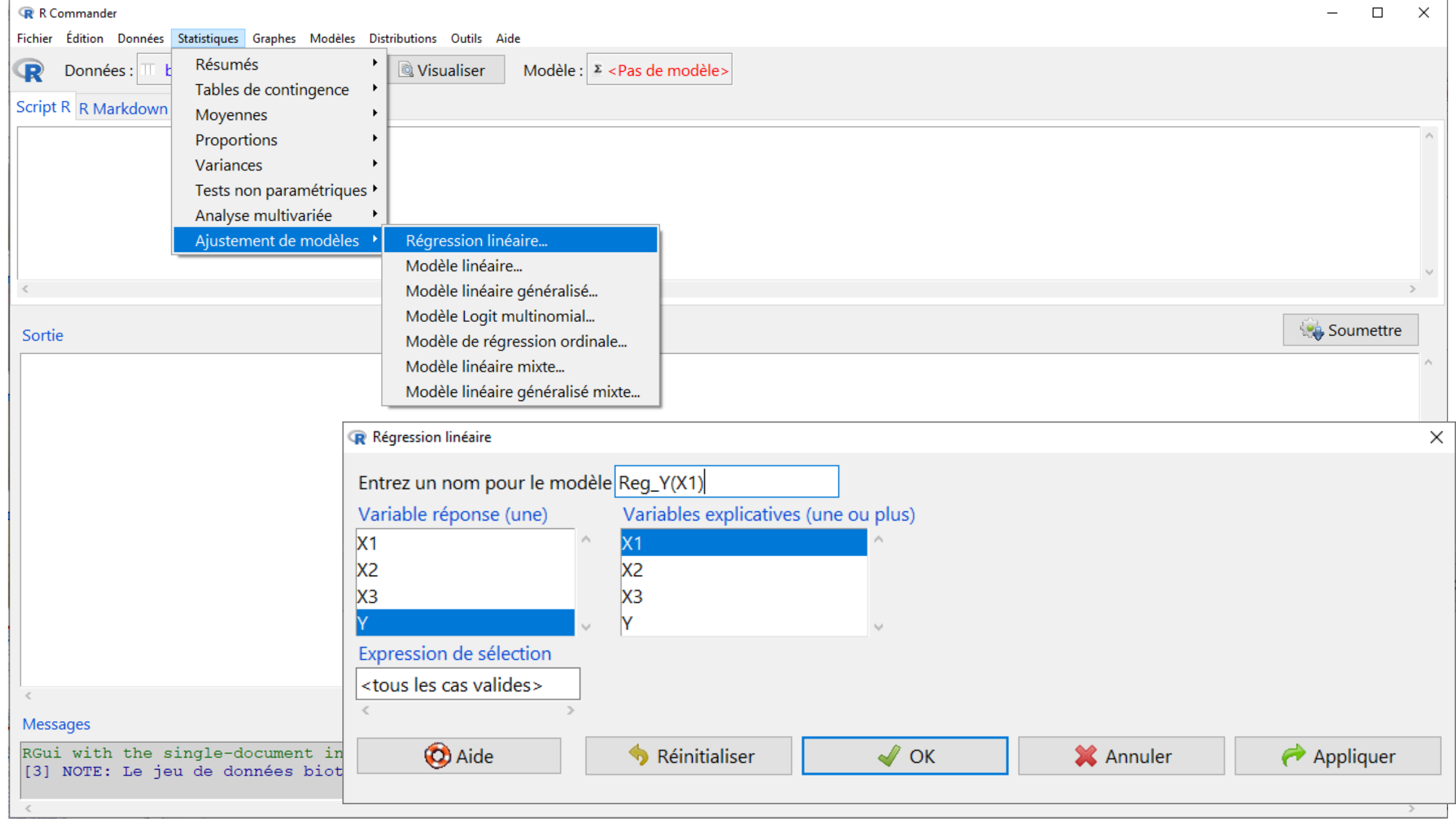
Taille du texte des libellés d'axes : 1.0

Position de la légende

- ◉ Au dessus de graphe
- ○ En haut à gauche
- ○ En haut à droite
- ○ En bas à gauche
- ○ En bas à droite

Aide | Réinitialiser | OK | Annuler




R Commander
Fichier Édition Données Statistiques Graphes Modèles Distributions Outils Aide
Résumés
Tables de contingence
Moyennes
Proportions
Variances
Tests non paramétriques
Analyse multivariée
Ajustement de modèles
Régression linéaire...
Modèle linéaire...
Modèle linéaire généralisé...
Modèle Logit multinomial...
Modèle de régression ordinale...
Modèle linéaire mixte...
Modèle linéaire généralisé mixte...
Données :
Visualiser
Modèle : <Pas de modèle>
Script R
R Markdown
Sortie
Soumettre
Messages
RGui with the single-document in
[3] NOTE: Le jeu de données biot
Régression linéaire
Entrez un nom pour le modèle
Reg_Y(X1)
Variable réponse (une)
X1
X2
X3
Y
Variables explicatives (une ou plus)
X1
X2
X3
Y
Expression de sélection
<tous les cas valides>
Aide
Réinitialiser
OK
Annuler
Appliquer

R Commander

Fichier Édition Données Statistiques Graphes Modèles Distributions Outils Aide

Données : biotope | Éditer | Visualiser | Modèle : RegModel.2

Script R | R Markdown

```
RegModel.2 <- lm(Y~X1, data=biotope)
summary(RegModel.2)
```

Soumettre

Sortie

```
> summary(RegModel.2)

Call:
lm(formula = Y ~ X1, data = biotope)

Residuals:
    Min      1Q  Median      3Q     Max
-8.3860 -1.9545 -0.1913  2.0671  7.2124

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 26.032594   0.457843   56.86   <2e-16 ***
X1           0.047537   0.002691   17.67   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.259 on 198 degrees of freedom
Multiple R-squared:  0.6119, Adjusted R-squared:  0.6099
F-statistic: 312.1 on 1 and 198 DF,  p-value: < 2.2e-16
```

Messages

```
[3] NOTE: Le jeu de données biotope a 200 lignes et 5 colonnes.
[4] ERREUR: "Reg_Y(X1)" n'est pas un nom correct.
```

Régression linéaire : résultats

```
> summary(RegModel)

Call:
lm(formula = Y ~ X1, data = biotope)

Residuals:
    Min      1Q  Median      3Q     Max
-8.3860 -1.9545 -0.1913  2.0671  7.2124

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 26.032594   0.457843   56.86   <2e-16 ***
X1           0.047537   0.002691   17.67   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.259 on 198 degrees of freedom
Multiple R-squared:  0.6119,   Adjusted R-squared:  0.6099
F-statistic: 312.1 on 1 and 198 DF,  p-value: < 2.2e-16
```

Régression linéaire : résultats

Régression linéaire simple

Modèle (> population) : $Y = \beta_0 + \beta_1 X + \varepsilon$

$Y = b_0 + b_1 X + E$

Y : VD, VA quantitative continue (unités non précisées) > mesures biotope

X : Variable fixée quantitative continue (donc, à priori ce n'est pas une VA)

$Y = 26,032594 + 0,0047537.X$ *>>>> Cette équation est-elle bien réaliste?*

$b_0 = 26,032594$ ordonnée à l'origine

$b_1 = 0,0047537$ pente de la droite de régression

Régression linéaire : résultats

Régression linéaire simple

$Y = 26,032594 + 0,0047537.X$ >>>> *Cette équation est-elle bien réaliste?*

Méthode des moindres carrés ordinaires (MCO, OLS in english)

Les erreurs sont (globalement) minimisées :
méthode mathématique basée sur la dérivation (équation matricielle)
+ utilisation de la densité de probabilité de la Loi Normale

Permet d'obtenir (à la main, avec une machine à calculer, avec R, Python ou tout autre logiciel) :

$b_0 = 26,032594$ ordonnée à l'origine

$b_1 = 0,0047537$ pente de la droite de régression