Statistiques descriptives

François Simonneau

IUT Nantes

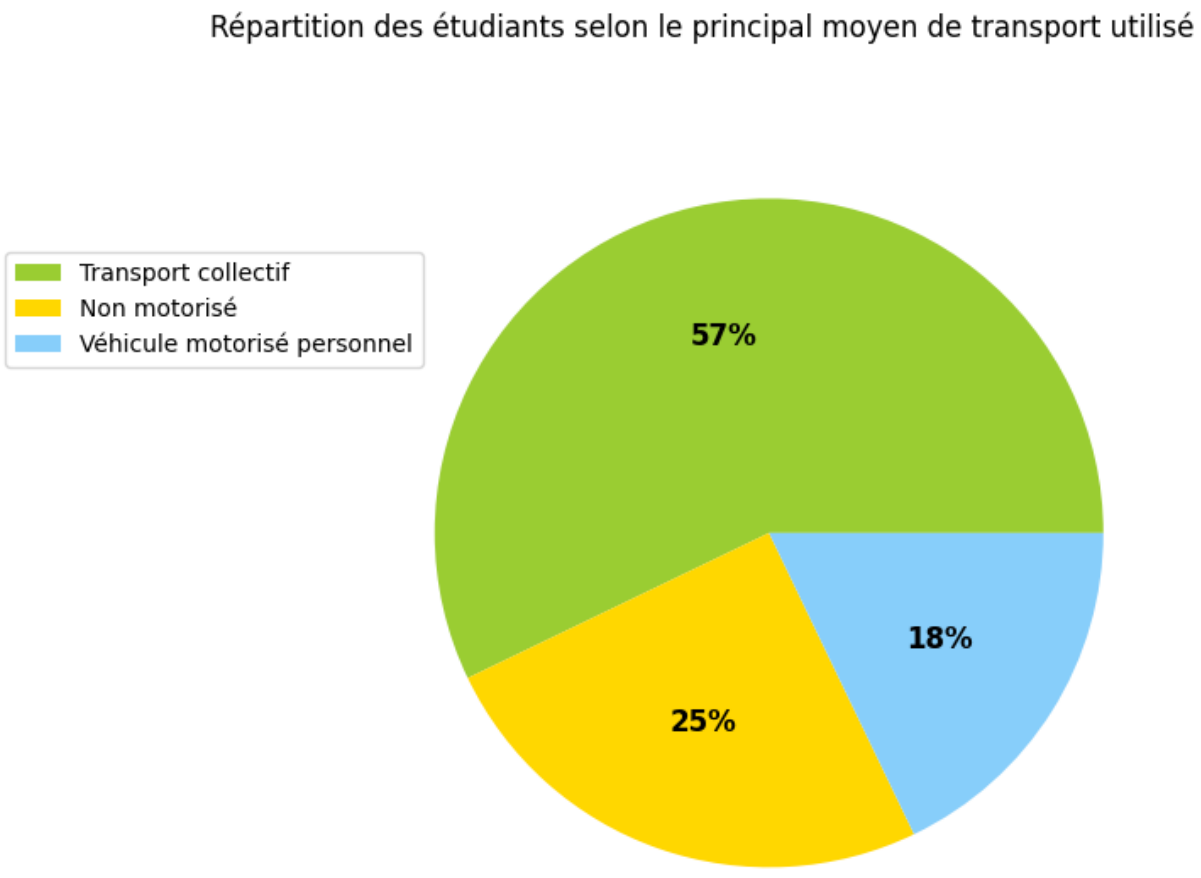
1. Différents types de variables
2. Différents indicateurs
3. Méthode de régression linéaire

Exemple étude

id_etu	Transport principal	Temps trajet	Nb moyens transport employés
1	Transport collectif	57	2
2	non motorise	39	2
3	vehicule motorise personnel	30	1
4	non motorise	3	1
5	Transport collectif	72	2
6	Transport collectif	92	4
7	Transport collectif	15	2
8	non motorise	15	1
9	non motorise	45	1
10	Transport collectif	69	2
11	Transport collectif	23	2
12	vehicule motorise personnel	107	1
13	non motorise	12	1
14	Transport collectif	20	3
15	Transport collectif	137	5
16	Transport collectif	21	2
17	Transport collectif	24	2
18	non motorise	51	2
19	Transport collectif	14	1
20	Transport collectif	54	3
21	vehicule motorise personnel	10	1

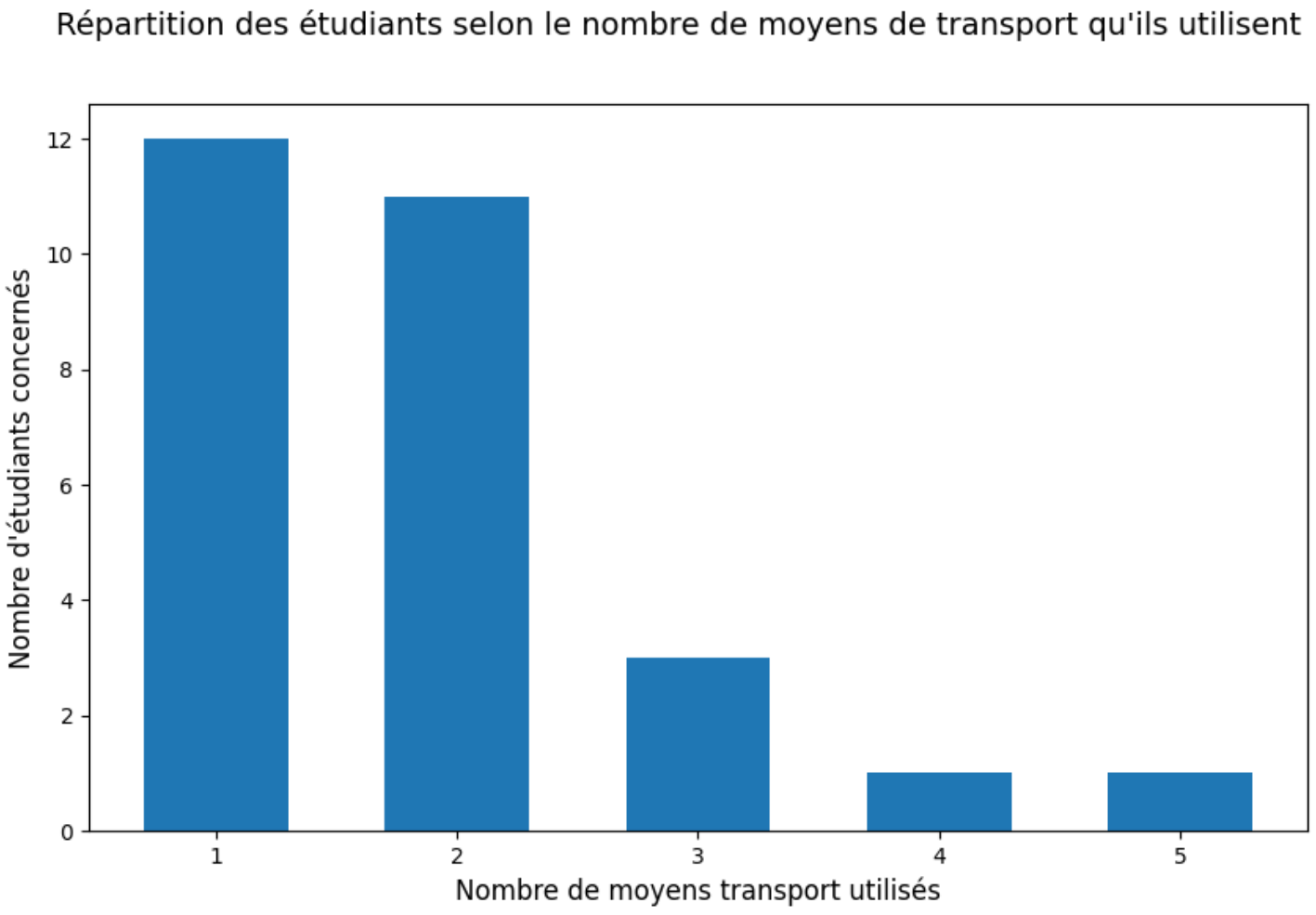
Différents types de variables

Qualitatives



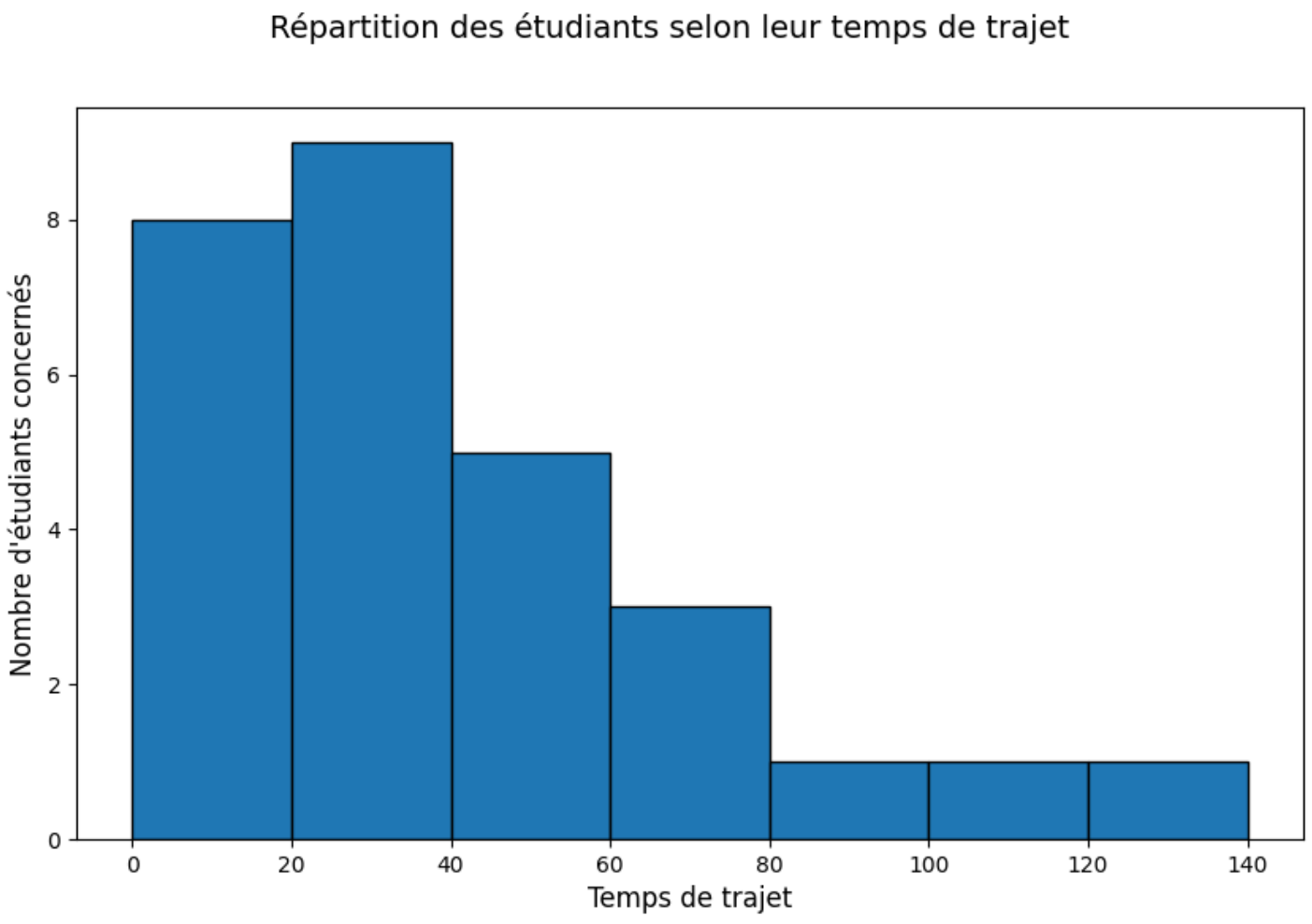
Différents types de variables

Quantitatives discrètes



Différents types de variables

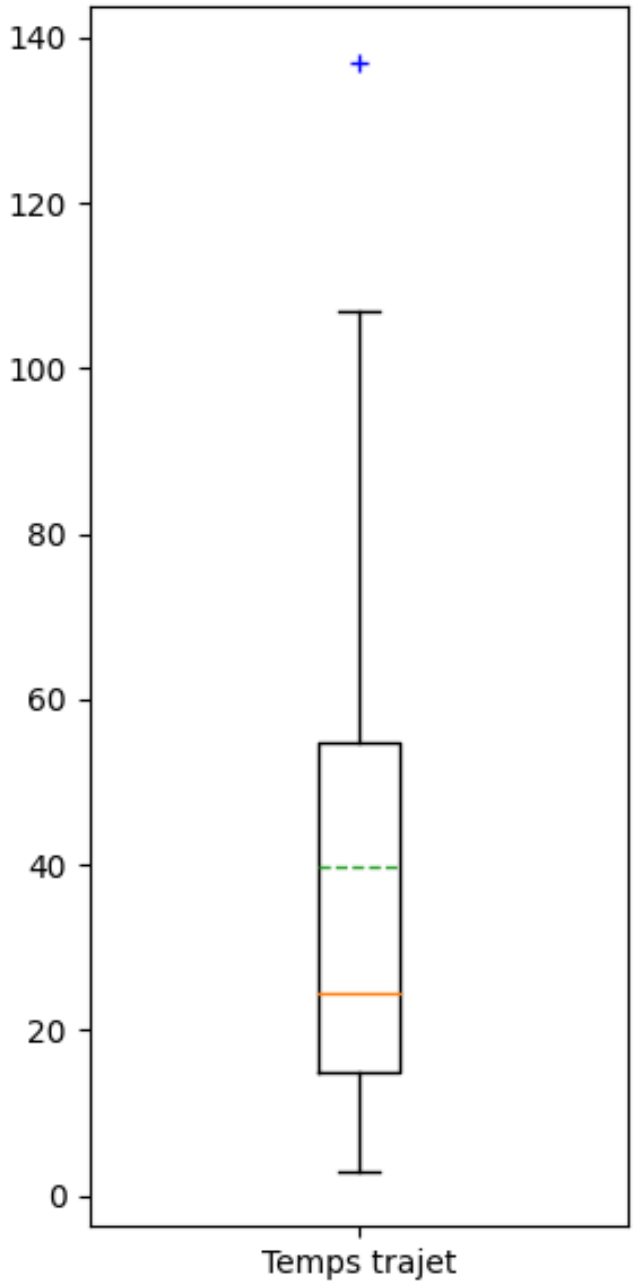
Quantitatives continues



Indicateurs de position

- Moyenne (sensibilité aux valeurs extrêmes)
- Médiane (Attention au sens quand nb modalités faible ou avec certaines modalités très représentées)
- Quartiles
- Déciles

count	28.000000
mean	39.714286
std	32.535857
min	3.000000
25%	15.000000
50%	24.500000
75%	54.750000
max	137.000000



Indicateurs de dispersion

Selon les écarts entre les observations et la moyenne :

- Variance
- Ecart type

Selon les écarts ou rapports entre indicateurs de positions :

- Ecart inter-quartiles (Q3-Q1)
- Ecart inter-déciles (D9-D1)
- Rapport inter-quartiles (Q3/Q1)
- Rapport inter-déciles (D9/D1)

Variance, écart type

						Moyennes
Tps x_i (mn)	2	8	10	15	25	

Variance, écart type

						Moyennes
Tps x_i (mn)	2	8	10	15	25	$m = 12$ (mn)

Variance, écart type

						Moyennes
Tps x_i (mn)	2	8	10	15	25	$m = 12$ (mn)
$x_i - m$	-10	-4	-2	3	13	

Variance, écart type

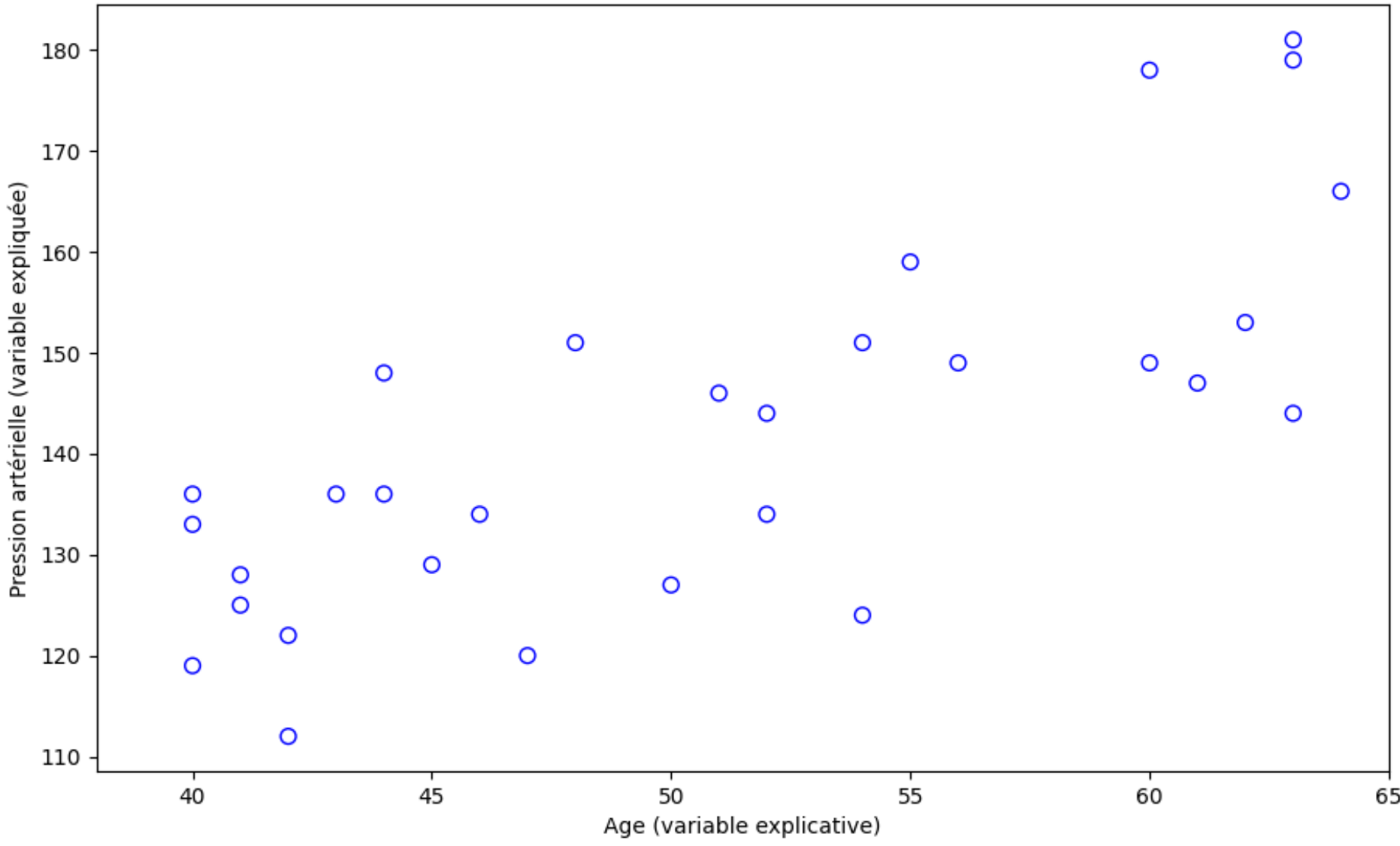
						Moyennes
Tps x_i (mn)	2	8	10	15	25	$m = 12$ (mn)
$x_i - m$	-10	-4	-2	3	13	0

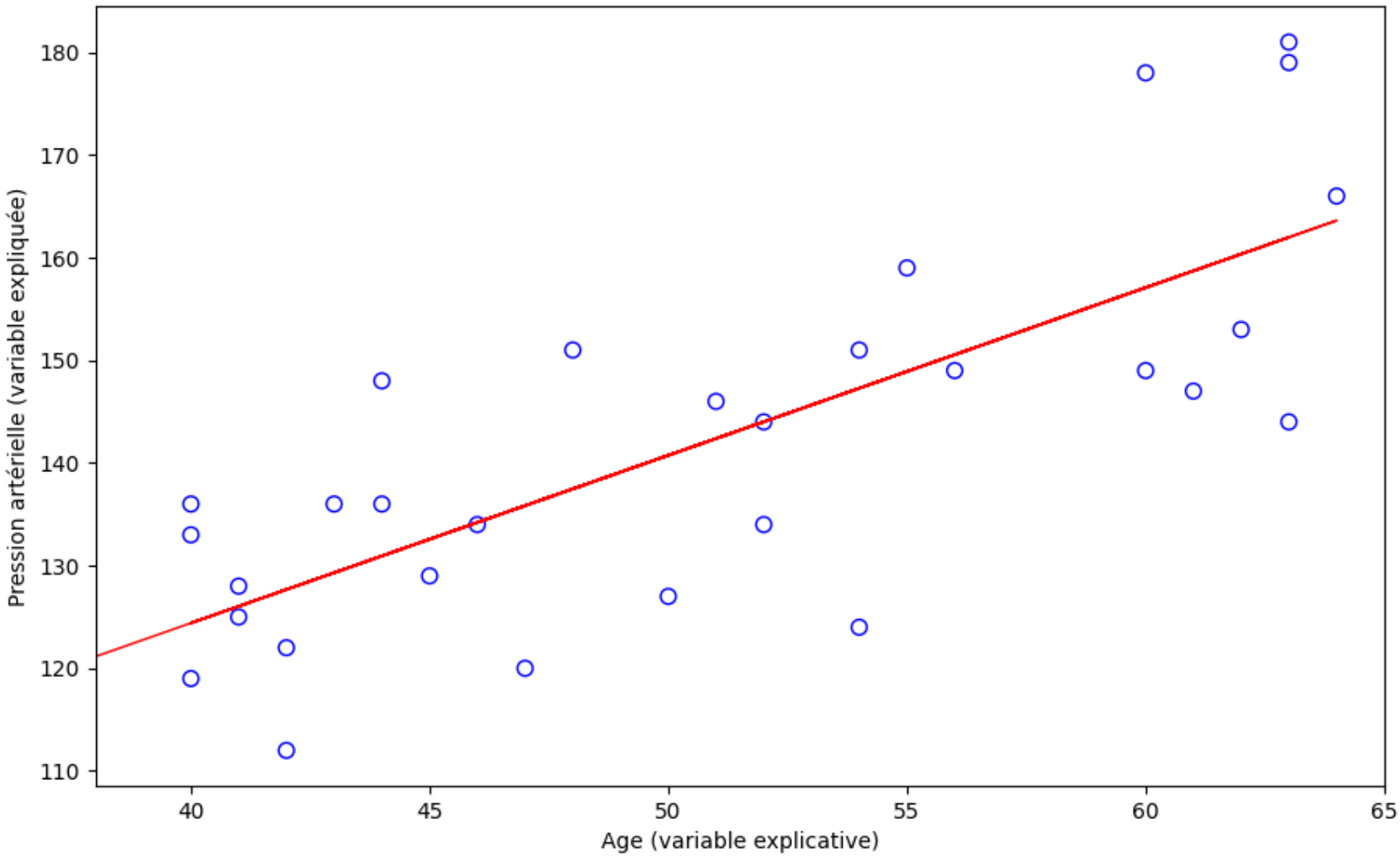
Variance, écart type

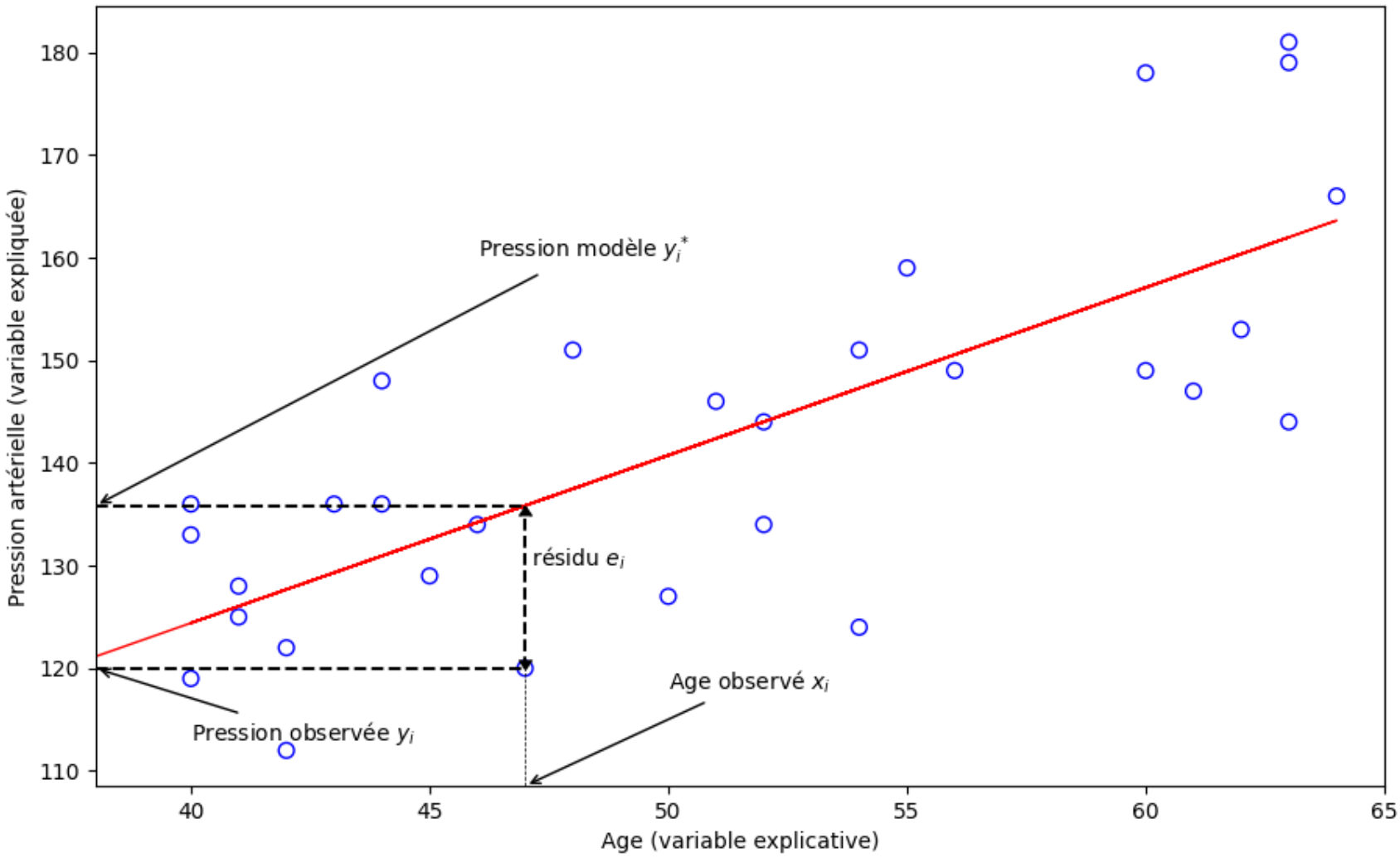
						Moyennes
Tps x_i (mn)	2	8	10	15	25	$m = 12$ (mn)
$x_i - m$	-10	-4	-2	3	13	0
$(x_i - m)^2$	100	16	4	9	169	$Var = 59,6\ mn^2$

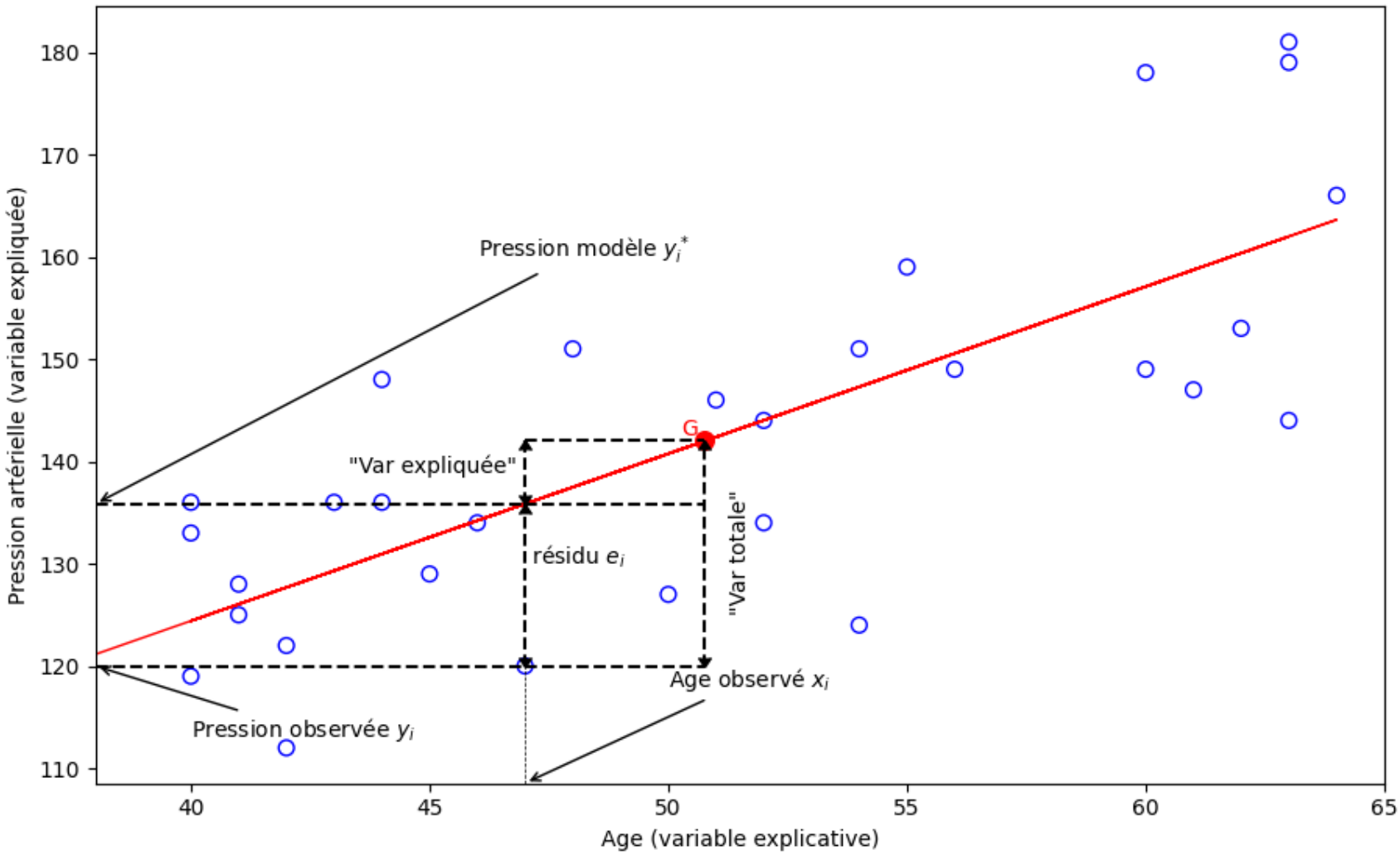
On définit alors l'écart type comme la racine carrée de la variance et qui peut s'exprimer avec la même unité que la variable étudiée :

$$\sigma = \sqrt{Var} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - m)^2} \simeq 7,7\ mn$$


Pression artérielle (variable expliquée)
180
170
160
150
140
130
120
110
40
45
50
55
60
65
Age (variable explicative)


Pression artérielle (variable expliquée)
Age (variable explicative)
180
170
160
150
140
130
120
110
40
45
50
55
60
65


Pression artérielle (variable expliquée)
Age (variable explicative)
Pression modèle y_i^*
résidu e_i
Age observé x_i
Pression observée y_i
180
170
160
150
140
130
120
110
40
45
50
55
60
65


Pression artérielle (variable expliquée)
Age (variable explicative)
Pression modèle y_i^*
G
"Var expliquée"
résidu e_i
"Var totale"
Age observé x_i
Pression observée y_i
180
170
160
150
140
130
120
110
40
45
50
55
60
65

Méthode des moindres carrés

La méthode consiste à rechercher la droite qui passera au plus près de l'ensemble des points en prenant comme critère le fait que la somme des carrés des résidus pour l'ensemble des observations soit la plus petite possible. En considérant l'équation d'une droite selon l'écriture suivante : $y = ax + b$, chaque résidu s'écrit $e_i = y_i - ax_i - b$ et le problème se résume au fait de chercher les valeurs de a et b qui vont minimiser la fonction de deux variables suivante :

$$M(a, b) = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - ax_i - b)^2$$

Theorem 1 (Théorème)

Les coefficients a et b qui minimisent le critère des moindres carrés sont donnés par :

$$a = \frac{Cov(X,Y)}{\sigma_X^2} \text{ et } b = \bar{y} - a\bar{x}$$

où $Cov(X,Y)$ désigne la covariance des variables X et Y et est définie par :

$$Cov(X,Y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

Le min s'obtient en annulant les dérivées partielles par rapport à a et b :

$$\begin{cases} M'_a(a,b) &= -\sum_{i=1}^{n} 2(y_i - ax_i - b)x_i = 0 \\ M'_b(a,b) &= -\sum_{i=1}^{n} 2(y_i - ax_i - b) = 0 \end{cases}$$

En résolvant ce système de deux équations à deux inconnues on obtient la solution énoncée. De la deuxième équation découle notamment la relation $b = \bar{y} - a\bar{x}$ qui nous informe en particulier du fait que la droite de régression passe par le point $G(\bar{x}, \bar{y})$

Décomposition de la variance

En décomposant les écarts à la moyenne de la manière suivante :

$$y_i - \bar{y} = y_i - y_i^* + y_i^* - \bar{y}$$

On peut démontrer la décomposition suivante (pas évidente du tout) :

$$\underbrace{\frac{1}{n}\sum(y_i - \bar{y})^2}_{Var_{tot}} = \underbrace{\frac{1}{n}\sum(y_i - y_i^*)^2}_{Var_{residuelle}} + \underbrace{\frac{1}{n}\sum(y_i^* - \bar{y})^2}_{Var_{expliquee}}$$

On définit alors le coefficient de détermination r^2 comme la part de variance expliquée par le modèle et on montre que le coefficient de corrélation r est obtenu :

$$r^2 = \frac{Var_{exp}}{Var_{tot}} \text{ et } r = \frac{Cov(X,Y)}{\sigma_X \sigma_Y}$$

Résultats

Sur l'exemple étudié la méthode renvoie le modèle proposé suivant :

$$Tension = 1.63 \times Age + 59 \text{ et } r = 0.759$$

Le coefficient de corrélation étant plutôt proche de 1 il peut témoigner d'un lien linéaire plutôt probant entre les deux variables (mais on ne démontre pas ici qu'il ne pourrait pas y avoir un meilleur modèle). Le fait que le coefficient soit positif témoigne du fait que le lien est croissant pour la variable expliquée en fonction de la variable explicative.
Le coefficient directeur 1.63 témoigne de la tendance à avoir sur cet intervalle d'âge un "gain" de 1.63 point de tension par année qui passe.
Il n'est généralement pas souhaitable d'interpréter la valeur de l'ordonnée à l'origine qui témoigne de ce qui serait attendu par le modèle pour une valeur de la variable explicative de 0 ce qui peut être très éloigné de l'intervalle d'étude et faire que le modèle ne soit plus du tout valable à cet emplacement.