PrivateTeacher

Cours Privés de Science

Régression Linéaire

2024.09.08

V 1.5

Introduction

Les droites de regression linéaires sont des modèles dont on se sert pour représenter la réalité.

En tant que modèle, une droite possède plusieurs avantages par rapport aux observations. Une droite en effet, permet de :

1) Résumer une série d'observations
2) Expliquer les observations.
3) Prédire les prochaines observations.

Un modèle cependant n'est pas la réalité mais seulement une approximation de celle-ci. Pour cette raison, une erreur est toujours associée aux droites de régression, comme nous allons le voir dans un instant.

La place du marché

Un marchand de pomme vend le fruit de sa récolte sur la place du marché.

Voici les prix affiché sur son étalage :

0.5 kilo de pomme coûte 1.5 CHF

1 kilo de pomme coûte 3 CHF

10 kilo de pomme coûte 30 CHF

On l'a compris, pour savoir quel est le prix des pommes, il suffit de multiplier le poids des pommes par 3.

Si x représente le poids des pommes, et y le prix, alors on peut écrire la relation sous forme d'une équation :

$$\boxed{y = 3x}$$

On dit alors que le prix des pommes est <u>proportionnel</u> au poids, car le rapport entre le prix et le poids est toujours le même quel que soit le nombre de pomme : $\frac{30}{10} = 3$ $\frac{9}{3} = 3$

On appel 3 le <u>facteur de proportionnalité</u>

En mathématiques, deux valeurs sont dites proportionnelles si leur rapport reste identique, même quand ces valeurs changent.

si	1	kg	coûte	3 CHF
alors	2	kg	coûtent	$3 \cdot 2 = 6$
et	4	kg	coûtent	$3 \cdot 4 = 12$
et	10	kg	coûtent	$3 \cdot 10 = 30$

$$3x = y$$

Cette relation simple permet de <u>résumer</u> toute la liste des prix.

Elle permet également d'<u>expliquer</u> le prix : si je paie cher, c'est par-ce que j'ai pris beaucoup de pommes.

Cette relation enfin permet de <u>prédire</u> un prix qui ne serait pas affiché par le marchand : si je choisit 3.45 kilos de pommes, alors je paie $3.45 \cdot 3 = 10.35$ CHF

Représentation graphique

La droite est une manière visuelle de représenter une relation de proportionnalité.

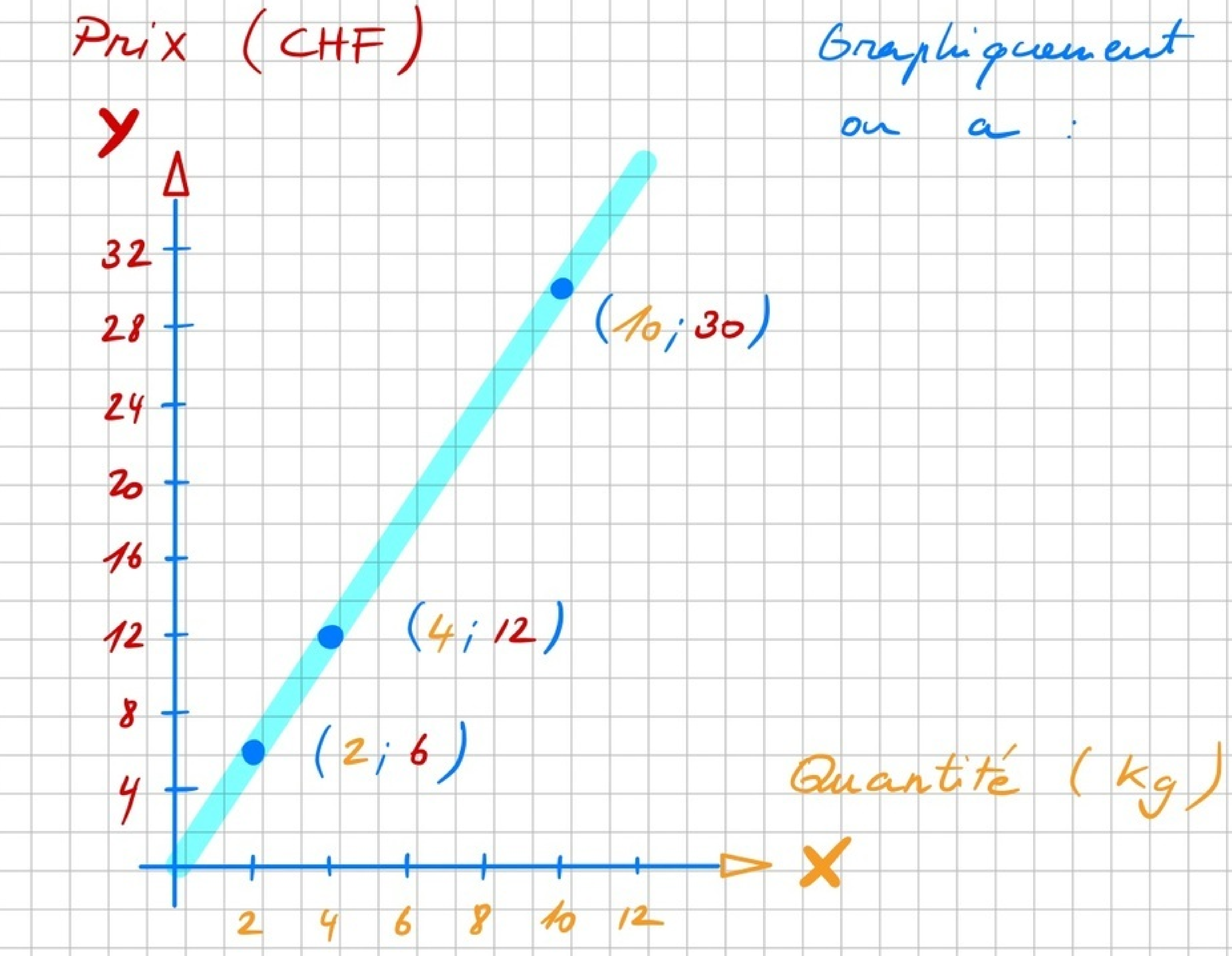
On la construit à l'aide d'un tableau :

x	y
2	6
4	12
10	30

Il existe une relation entre la quantité de pommes x et le prix des pommes y

Mathématiquement elle prend la forme $y = 3x$

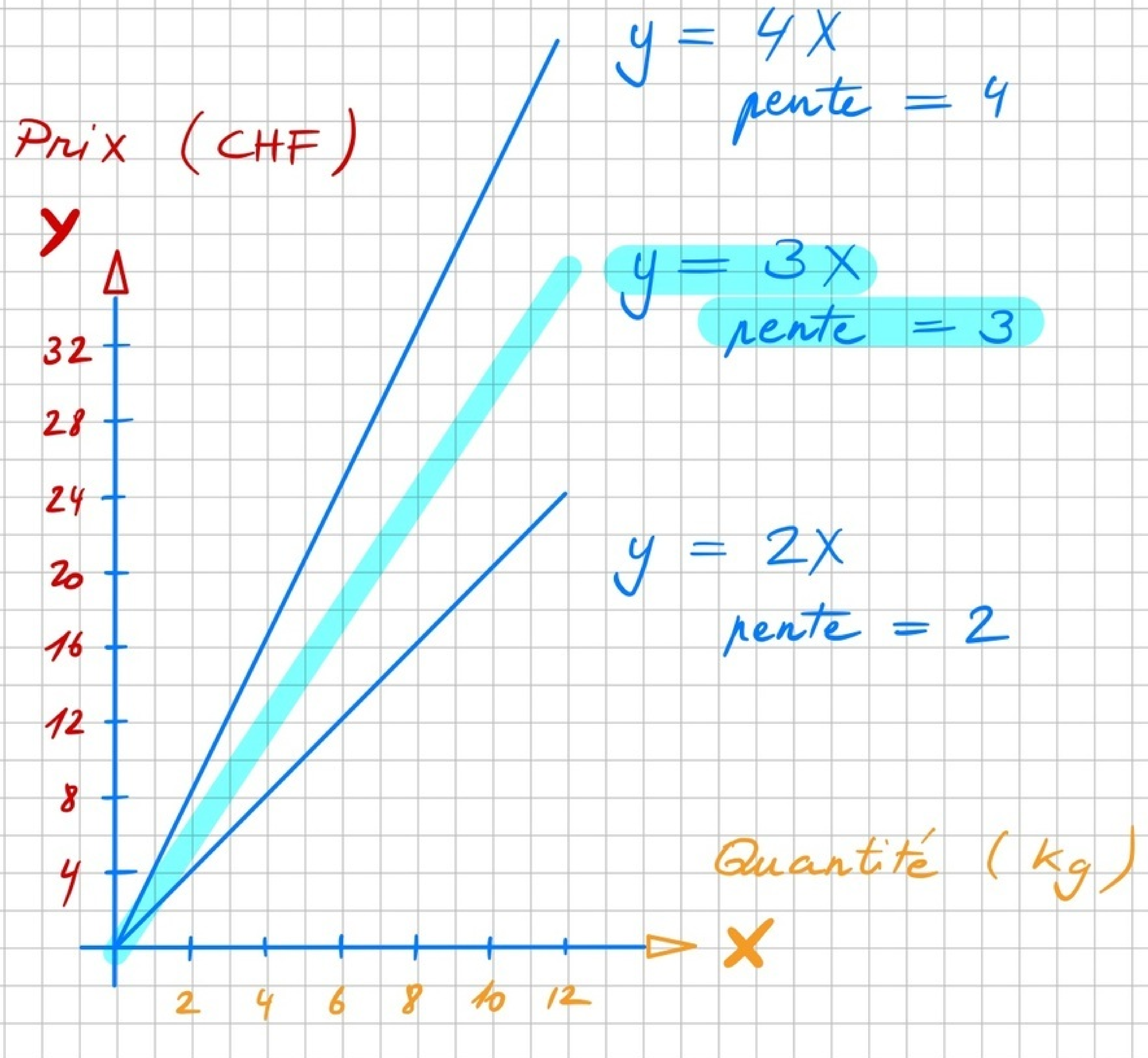
Graphiquement on a :



La pente

Si le prix du kilo de pomme avait été plus élevé, le prix des pomme augmenterait plus rapidement et la pente de la droite serai plus élevée

À l'inverse si le prix avait été plus bas, la pente aurait été plus faible



On peut lire la pente directement sur l'équation.

Elle est le facteur qui multiplie "x"

La pente représente le coefficient de proportionnalité entre le prix et la quantité.

À l'inverse, on peut lire la pente directement sur le graphique voici comment s'y prendre :

Relation de proportionnalité

En mathématiques, deux valeurs sont dites proportionnelles si leur rapport reste identique, même quand ces valeurs changent.

Prenons par exemple le changement de température avec l'altitude.

Il fait toujours plus froid lorsque l'on monte en altitude.

À chaque fois que l'on s'élève de 1 km (dénivellation de 1 km) la température diminue de 6 °C en moyenne.

Si l'on monte de 4 km, la température chute de 24 °C.

Lorsqu'on écrit le rapport entre la température et la dénivellation, on a :

$$\frac{6}{1} = 6 \qquad \frac{24}{4} = 6$$

Le rapport entre la température et la dénivellation est toujours le même

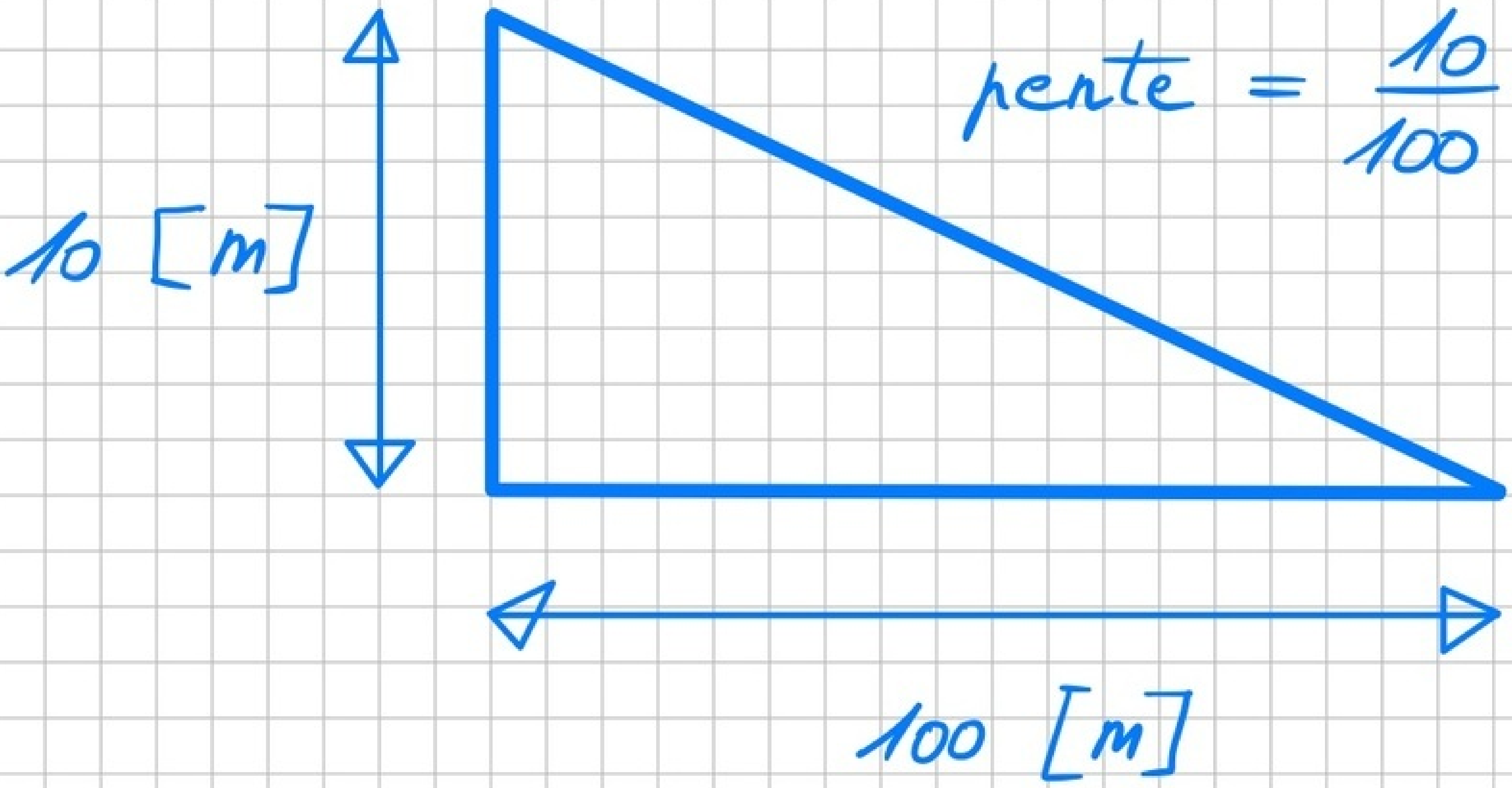
On dit alors que la température est proportionnelle à la dénivellation

Calcul de la pente

La pente indiquée sur ce panneau est de 10 %.

Cela signifie que pour 100 [m] parcouru <u>horizontallement</u>, la voiture sera descendue de 10 [m] <u>verticalement</u>.

Une pente est donc le rapport entre une distance verticale et une distance horizontale.



$$\text{pente} = \frac{10}{100} = 10\,\%$$

Ordonnée à l'origine

Imaginons maintenant que l'entrée à la place du marché soit payante.

Avant de pouvoir acheter des pommes au prix indiqué, il faut s'acquitter d'un montant de 8 CHF.

Le prix des pommes "y" est toujours proportionnel au nombre de pommes ($y = 3x$) mais cette fois, le montant de l'entrée s'y ajoute systematiquement si bien que le prix final devient :

$$y = 3x + 8$$

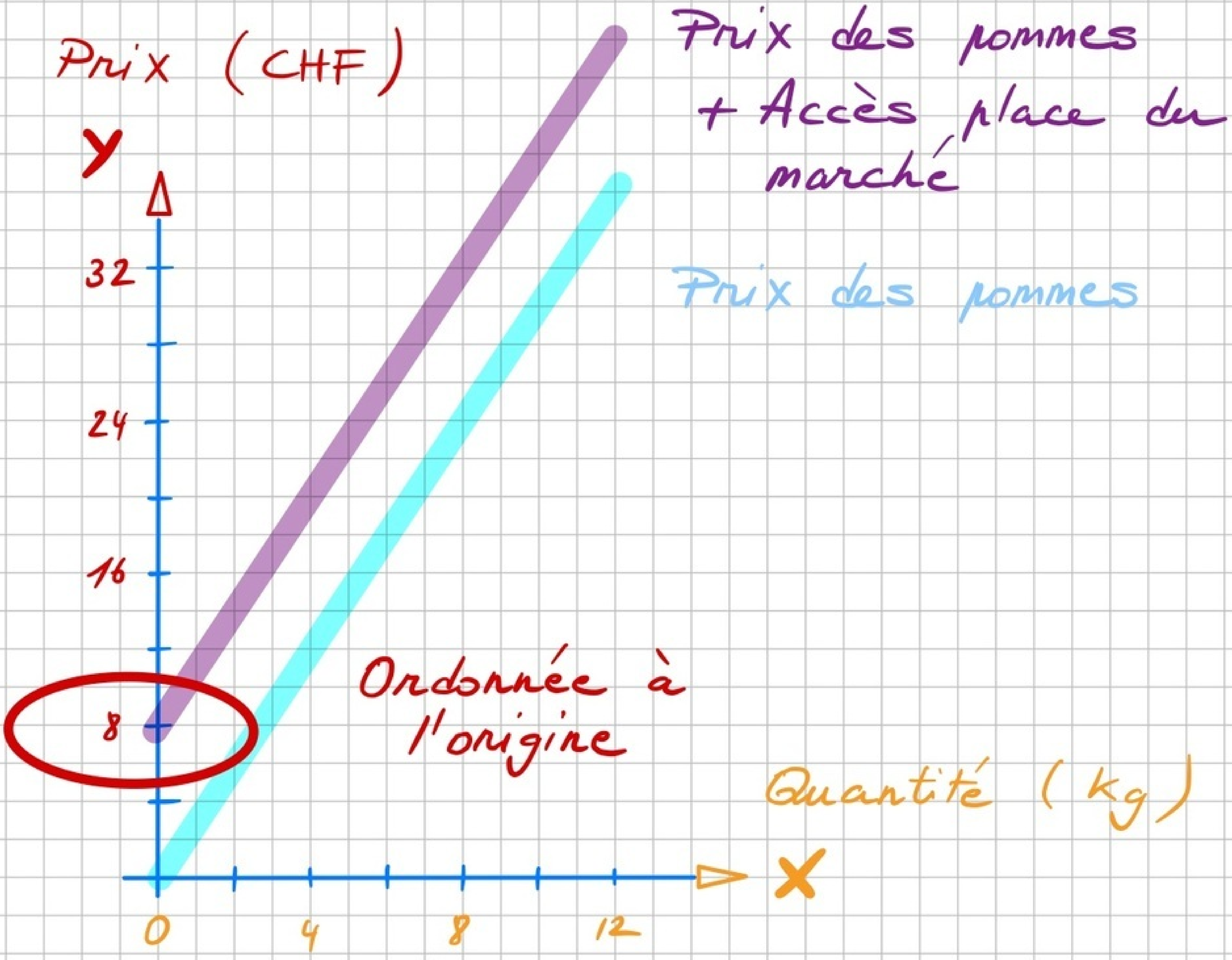
2 kg	coutent	$3 \cdot 2 + 8 = 14$
4 kg	coûtent	$3 \cdot 4 + 8 = 20$
10 kg	coûtent	$3 \cdot 10 + 8 = 38$

$$3x + 8 = y$$

Pour dessiner la relation entre le prix des pommes et le nombre de pommes, on refait notre tableau

x	y
2	14
4	20
10	38

Graphiquement, cela se traduit par une droite parallèle au dessus de la précédente :



On le voit, quelque soit le nombre de pommes, le prix final est toujours supérieur de 8 CHF.

Sur le graphique en $x = 0$ (0 pomme) on lit que le prix est de 8 CHF

Cette valeur de y en $x = 0$ se nomme <u>l'ordonnée à l'origine</u>

Equation de la droite

Les droites nous l'avons vu, sont déterminées par deux valeurs : sa pente et son ordonnée à l'origine.

Les droites <u>en générale</u> possède toutes une pente que l'on note "a" et une ordonnée à l'origine que l'on note "b"

$$\boxed{y = ax + b}$$

Il s'agit là de la <u>forme générale</u> de l'équation de la droite.

Cette forme désigne la famille de toutes les droites. L'équation :

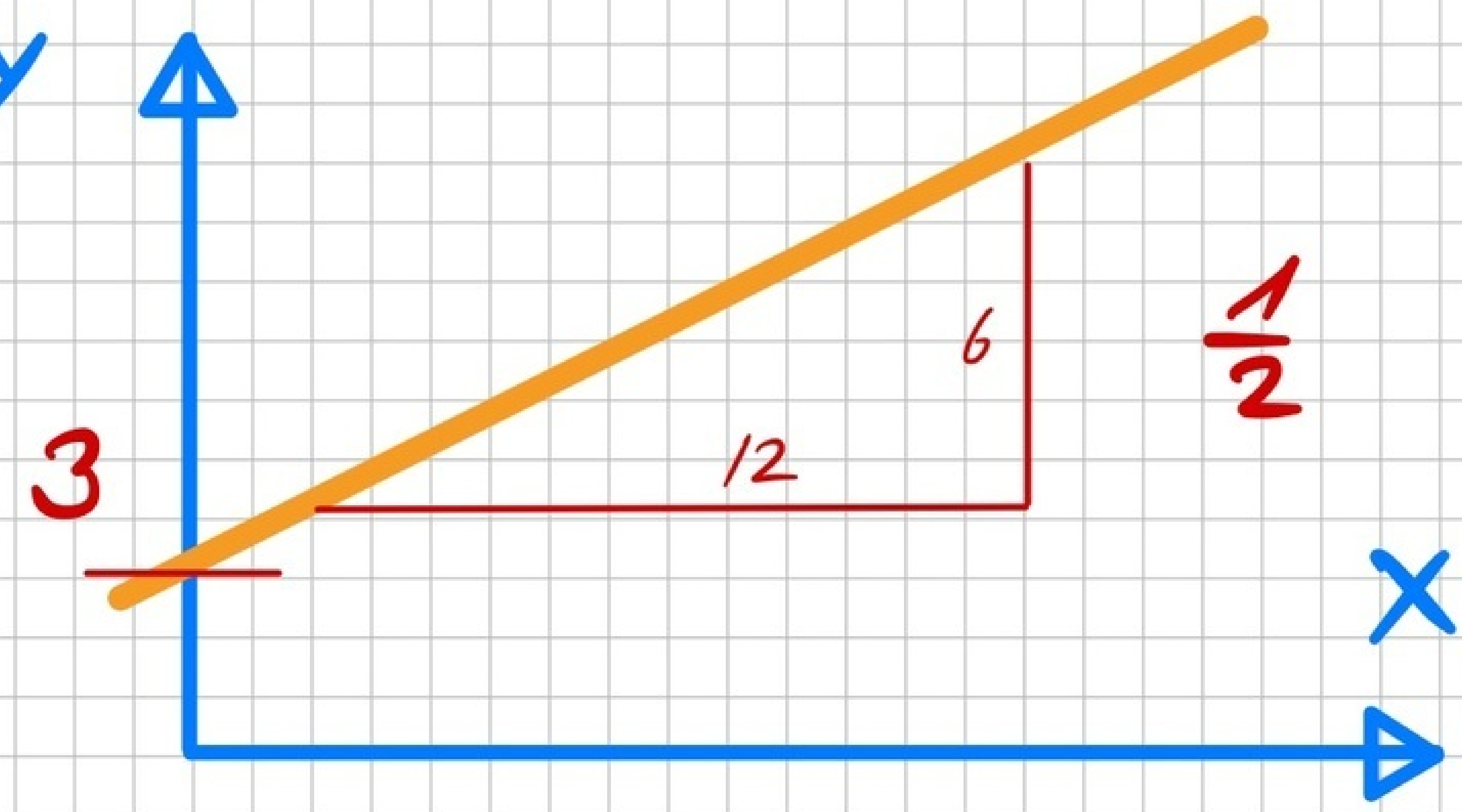
$$y = 4x + 2$$

désigne <u>une droite en particulier.</u>

On peut choisir "a" et "b" librement et obtenir n'importe quelle droite en particulier.

Pour cette raison, "a" et "b" se nomment <u>les paramètres de la droite.</u>

La droite en bref



On peut dessiner le graphique à partir de l'équation.

On peut lire l'équation sur le graphique

$$\boxed{y = \frac{1}{2}x + 3}$$

Régression linéaire

Lorsque l'on passe de l'équation générale de la droite $y = ax + b$ à une équation particulière p. exple $y = 4X + 7$, dans le but d'ajuster la droite à des observations on effectue ce qu'on appel une régression linéaire

La régression linéaire est un processus qui permet de trouver l'équation d'une droite particulière : la droite qui se superpose à nos observations.

De telles équations représentent donc un modèle de la réalité dont on peut dès lors se servir pour faire des prédictions.

Dans une certaine mesure on peut substituer le modèle à la réalité mais il ne faut jamais oublier que ceux-ci ne sont que des approximations

Nous avons vu quels étaient les avantages d'utiliser une droite pour représenter une relation de proportionnalité.

Voyons à présent comment trouver une telle relation au sein de nos données.

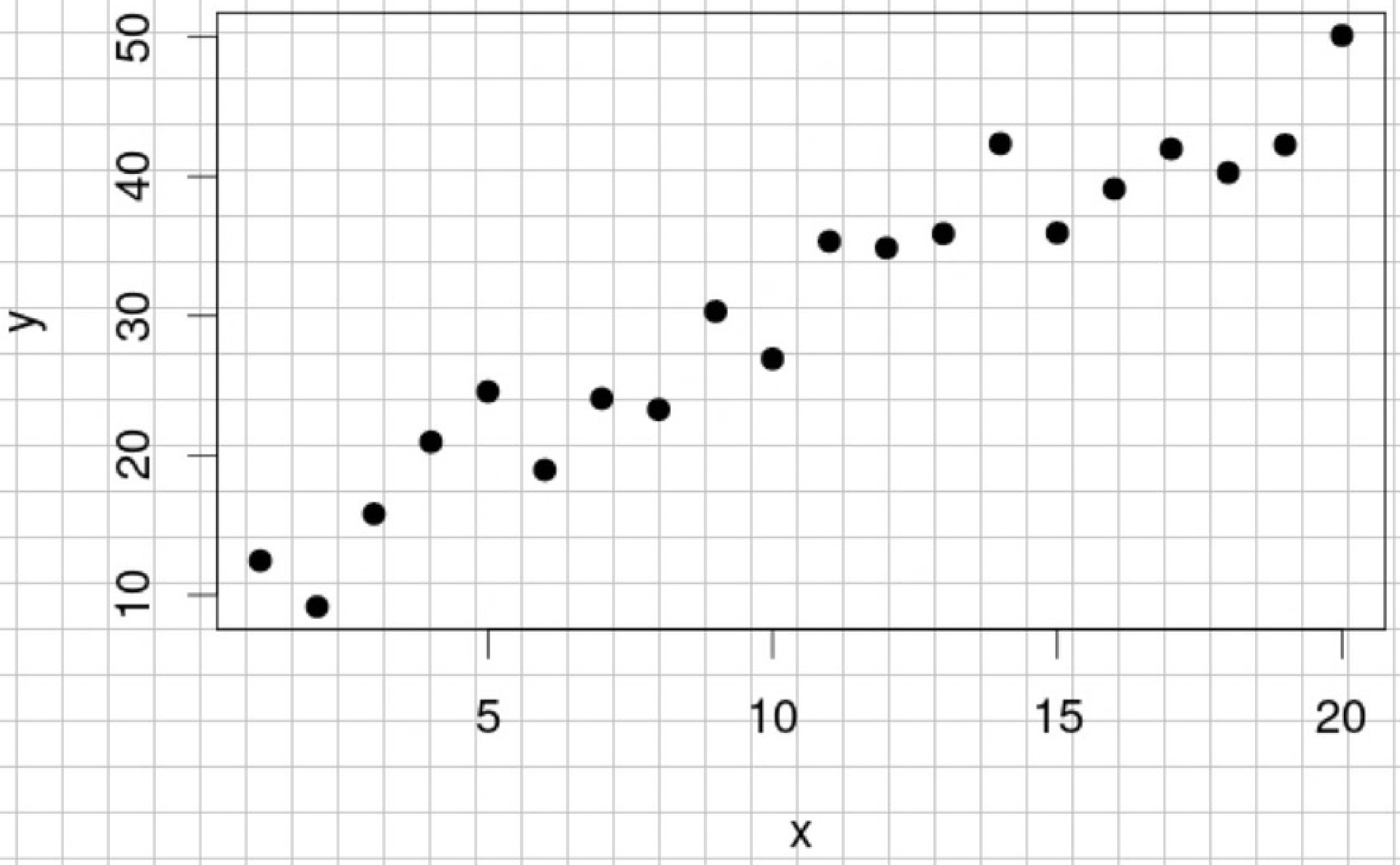
Nos observations sont réunies dans le tableau ci-contre.

Voici comment réaliser une régression linéaire à la main.

X	Y
1	12.45
2	9.14
3	15.80
4	20.96
5	24.57
6	18.95
7	24.07
8	23.28
9	30.31
10	26.92
11	35.34
12	34.86
13	35.89
14	42.34
15	35.95
16	39.10
17	41.97
18	40.26
19	42.26
20	50.09

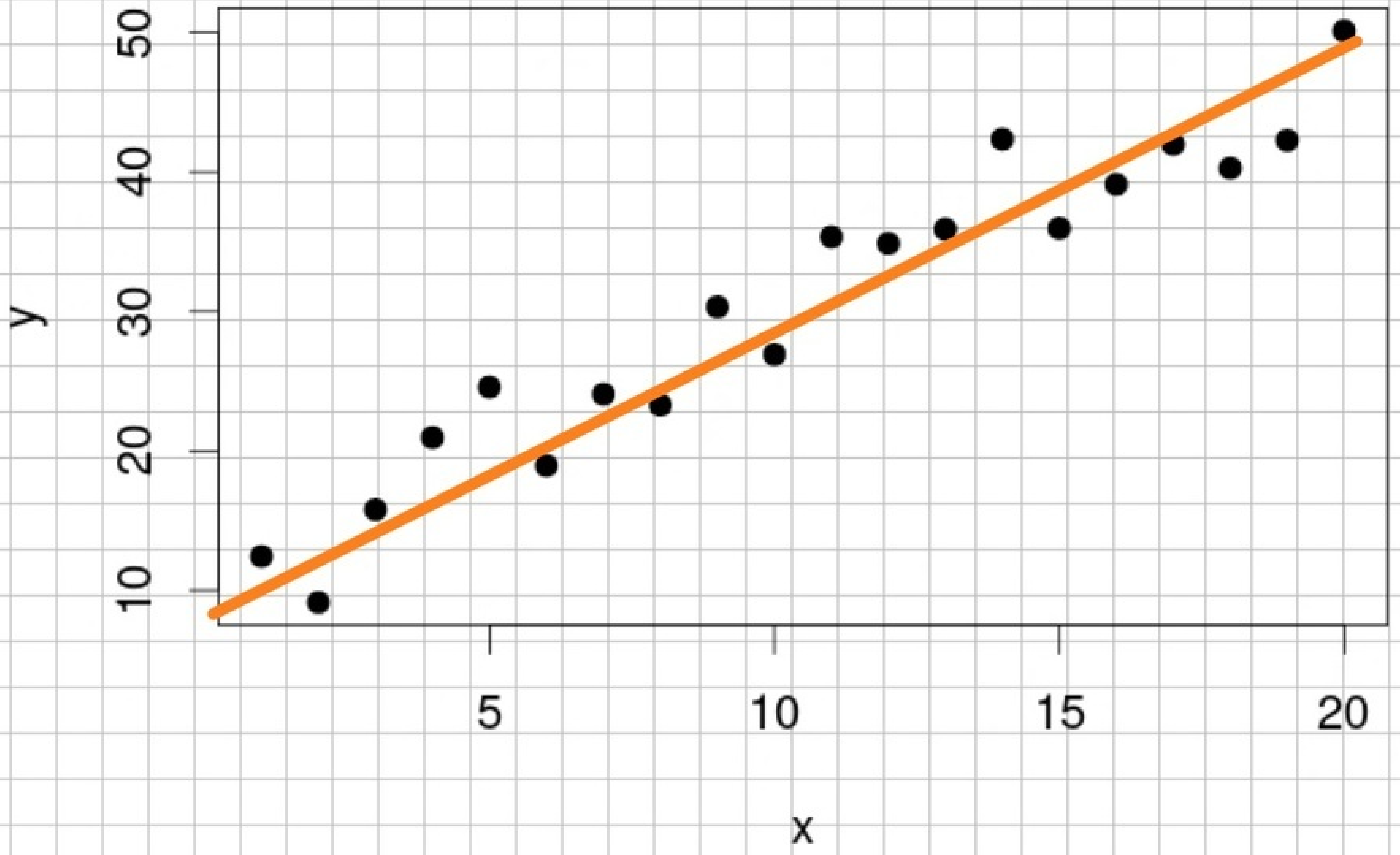
1) la première étape consiste à representer graphiquement une caractéristique X en fonction d'une caractéristique Y

On réalise cela à l'aide d'un <u>scatterplot</u>



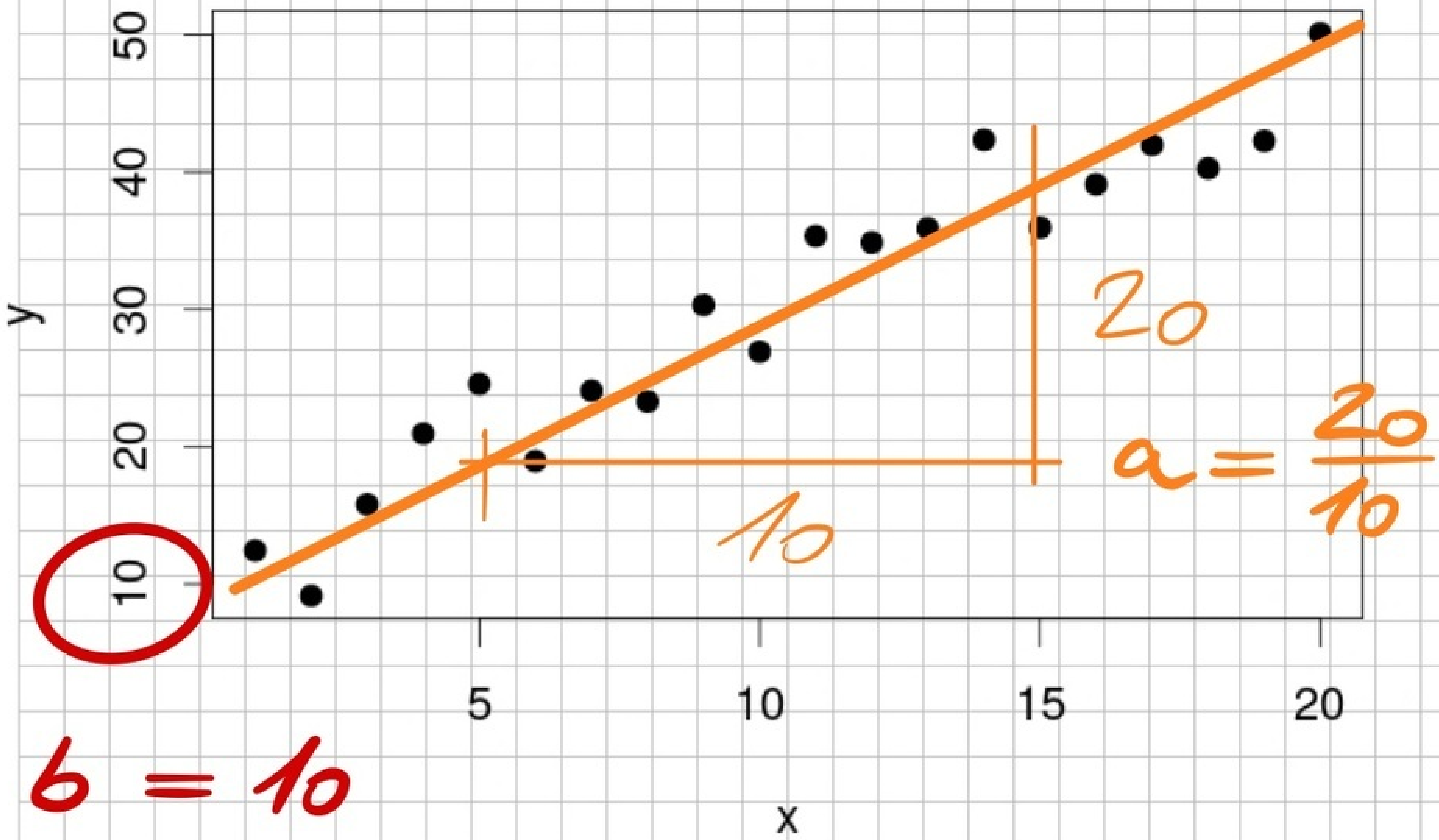
2) On ajuste ensuite une droite à la main. Le but du jeux est de trouver la droite qui passe <u>au plus près</u> de tous les points.

Par exemple celle-ci :



3) On utilise ensuite les outils vu auparavent pour déterminer l'équation de la droite :

$$y = aX + b$$



$a = \frac{20}{10} = 2$

$b = 10$

$$\boxed{y = 2X + 10}$$

Il s'agit là de notre modèle : l'équation de la droite de régression obtenue par régression linéaire sur nos observations.

Ce modèle $y = 2X + 10$ permet de :

Résumer l'information

À lui tout seul, ce modèle permet de résumer toutes les données. Plus besoin de 20 paires de points en effet, on sait qu'elles sont toutes situées autour de cette droite.

Prédire les valeurs

Grâce à ce modèle, on peut également prédire la valeur de y pour des X que l'on a pas du tout mesuré. $X = 40$ par exemple, n'apparaît même pas sur le graphique. À l'aide de ce modèle pourtant on peut calculer que la valeur y correspondante vaut $y = 2 \cdot 40 + 10 = 90$

Expliquer les variations de y

Ce modèle nous permet également d'expliquer les valeurs de y. Si y augmente par exple c'est par ce que X augmente. On peut même dire que y augmente deux fois plus que X car leur coefficient de proportionnalité vaut 2.

Voici à présent comment s'y prendrait une machine (un ordinateur) pour ajuster une droite sur cette même série de point.

Dans la pratique, un ordinateur utilise une méthode qui consiste a réduire au maximum la différence entre la droite et les points.

Voici comment cela fonctionne :

Une première droite est tracée approximativement.

Pour chaque point, un logiciel calcul à quelle distance il se trouve de la droite

Afin d'éviter que les distances positives ne compensent les distances négative, on les élève au carré

On fait ensuite la somme de chacune de ces distance au carré pour obtenir un score.

Plus le score est petit plus la droite est proche des points.

Le logiciel utilise ce score pour ajuster la droite.

La droite définitive est trouvée quand le score est minimum.

On appel cette méthode la méthode des moindres carré.

Voici comment se présente les résultats d'une régression linéaire faite à l'aide d'un logitiel de statistique.

Il s'agit d'une sortie logicielle (un output) du programme R. Il existe bien d'autres programmes capablent de faire la même chose mais dont les sorties logitielles sont différentes.

Tous cependant ont en commun un certain nombre de points essentiels, et ce sont ceux là que nous allons voir à présent.

```
Call:
lm(formula = y ~ x, data = df)

Residuals:
    Min      1Q  Median      3Q     Max
-5.4641 -2.4263 -0.2633  2.4766  5.6858

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   10.934      1.474   7.419 7.05e-07 ***
x              1.838      0.123  14.938 1.38e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.173 on 18 degrees of freedom
Multiple R-squared:  0.9254, Adjusted R-squared:  0.9212
F-statistic: 223.1 on 1 and 18 DF,  p-value: 1.382e-11
```

Reprenons cette sortie logicielle dans les détails
Elle se compose de 4 parties importante

1

```
Call:
lm(formula = y ~ x, data = df)
```

2

```
Residuals:
    Min      1Q  Median      3Q     Max
-5.4641 -2.4263 -0.2633  2.4766  5.6858
```

3

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   10.934      1.474   7.419 7.05e-07 ***
x              1.838      0.123  14.938 1.38e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

4

```
Residual standard error: 3.173 on 18 degrees of freedom
Multiple R-squared:  0.9254, Adjusted R-squared:  0.9212
F-statistic: 223.1 on 1 and 18 DF,  p-value: 1.382e-11
```

L'appel (call) de la fonction

```
Call:
lm(formula = y ~ x, data = df)
```

Cette partie est spécifique au logiciel R et n'a donc pas d'interêt si on utilise un autre logiciel.

Dans cette partie, le logiciel (R) nous dit simplement la commande qui lui a été donnée par l'utilisateur-trice.

On s'en sert pour vérifier que le logiciel est bien en train de faire ce qu'on pense qu'il fait ^^

Dans le cas présent, on a demandé à R de lire les données (data) dans une variable qui s'appel "df" pour en tirer un model linéaire (linear model, lm)

On lui précise ensuite que la variable dépendante (y) doit être expliquée à l'aide de la variable explicative x. $y \sim x$

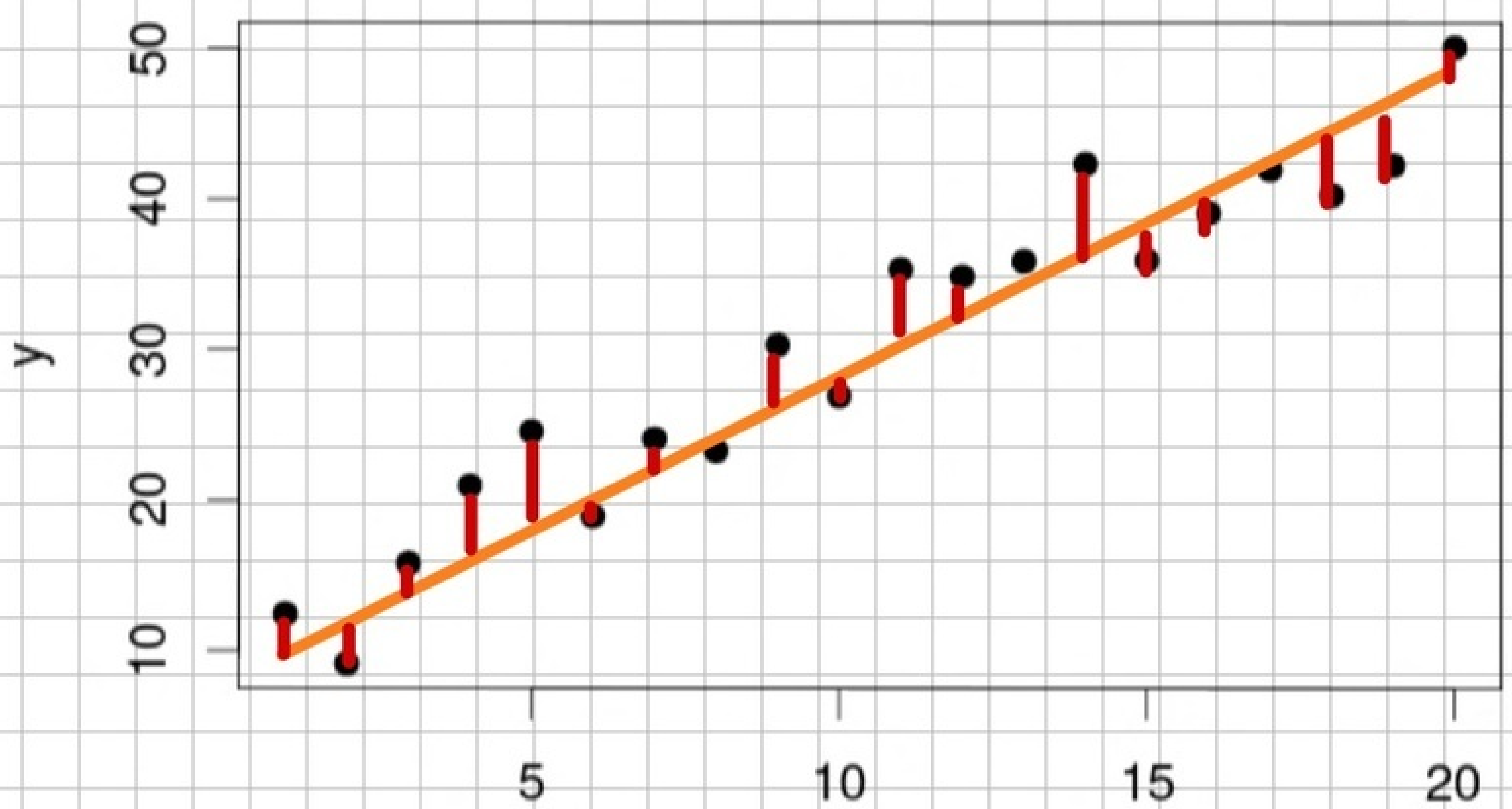
Le résidu du modèle

```
Residuals:
    Min      1Q  Median      3Q     Max
-5.4641 -2.4263 -0.2633  2.4766  5.6858
```

2

Une fois le processus de régression terminé et que la droite passant au plus près de tous les points à été trouvée on peut s'intéresser à la précision du modèle (par " modèle " on entend la droite de régression)

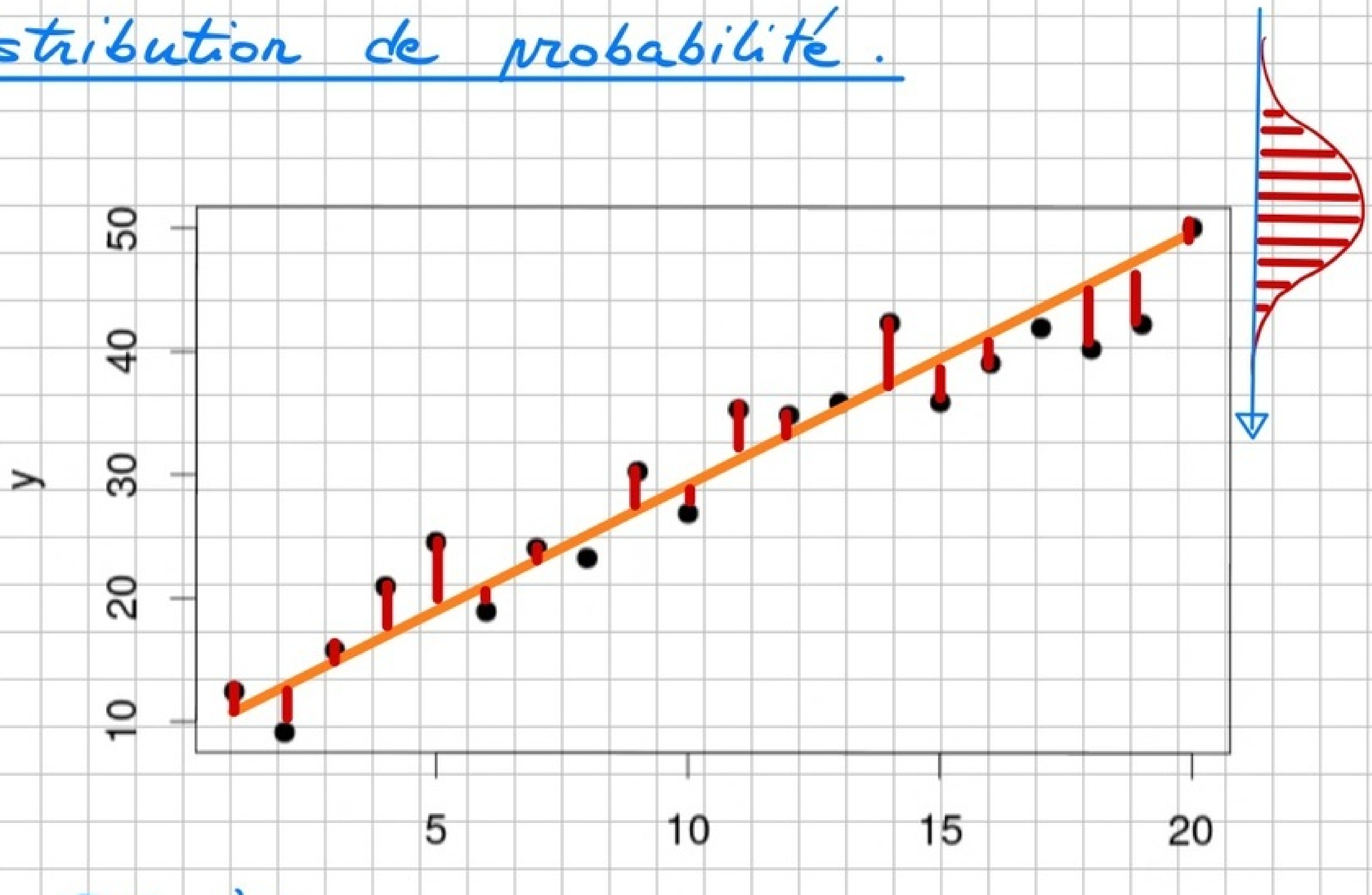
On dit que la droite a été ajustée au points. Or, quelque soit la qualité de cet ajustement, il demeure toujours une différence entre la droite et les points

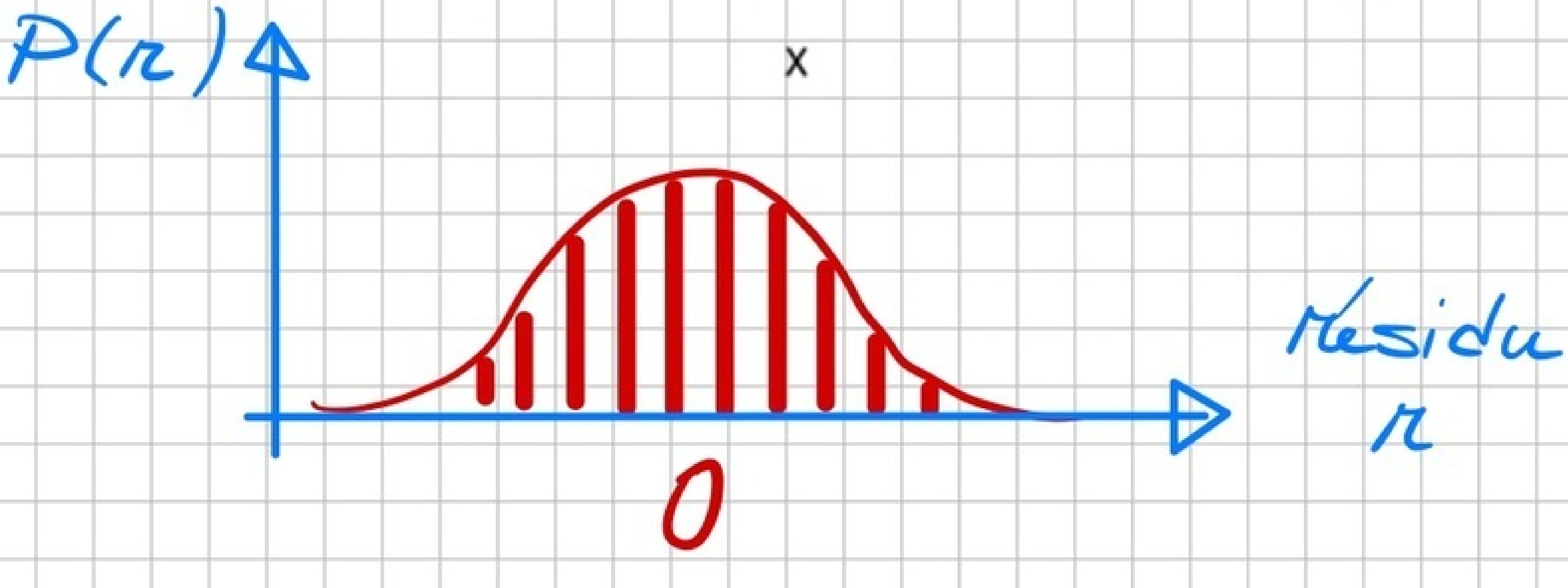


Si le rôle de la droite de régression est d'expliquer les variations de "y" à l'aide de "x", alors chacune des différences entre la droite et les points est une variation qui reste à expliquer

Pour cette raison on appel ces différences les résidus de la régression.

Il y a autant de résidus que de valeurs expérimentales. Tout comme ces valeurs, les résidus suivent une certaine distribution de probabilité.



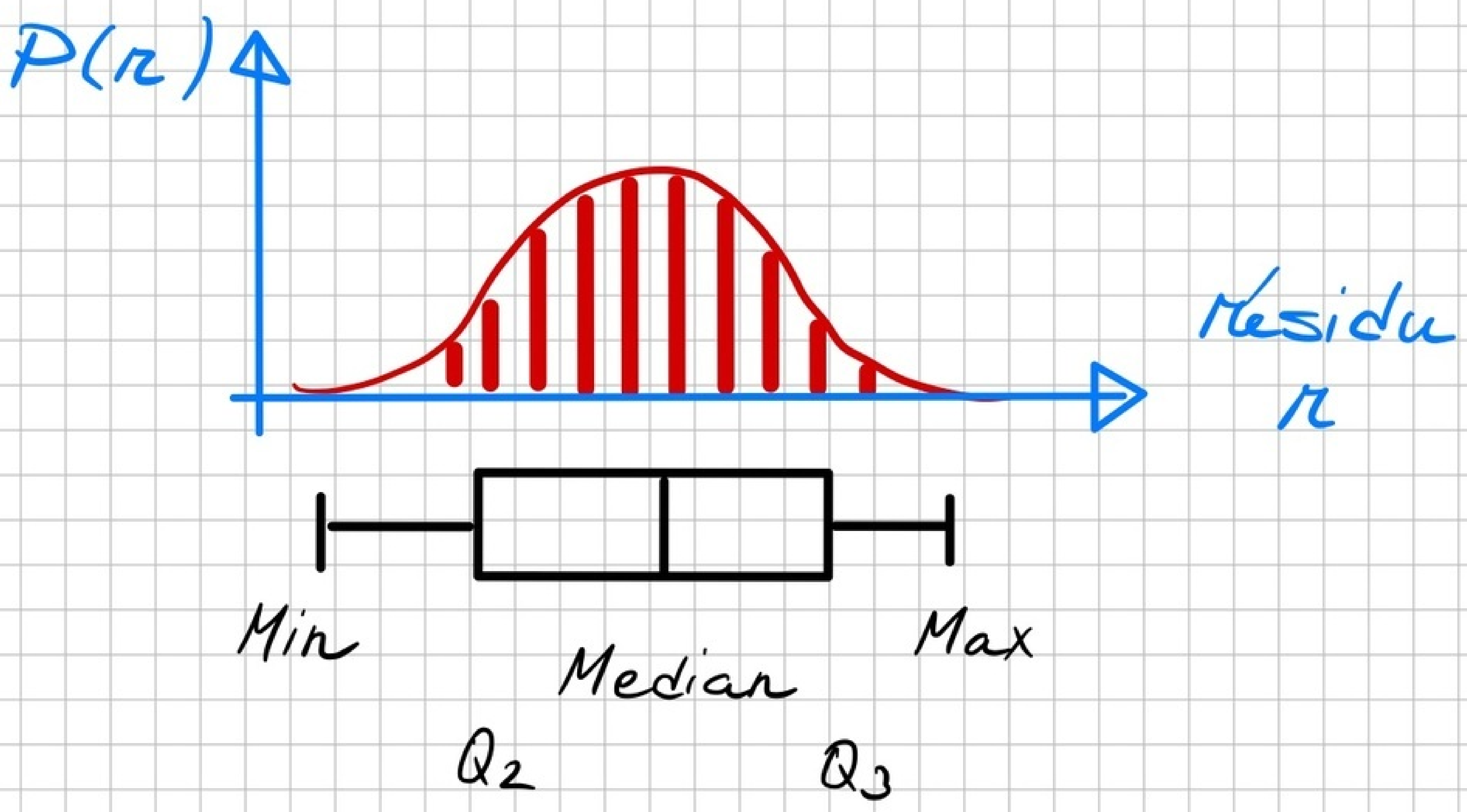


Lorsque le modèle est bon l'essentiel des variations sur "y" ont été expliquée Tout ce qui reste n'est que du bruit sans aucune valeur explicative.

Or, une caractéristique de ce "bruit" est qu'il se distribue normalement autours de zero.

Un résidu normallement (symétriquement) distribué autour de zero est donc une indication que le modèle explique les données correctement.

C'est la toute l'utilité d'afficher les différents quartiles de la distribution des résidus.



Residuals:

Min	1Q	Median	3Q	Max
-5.4641	-2.4263	-0.2633	2.4766	5.6858

2

On le voit, la distribution des résidus est centrée en 0 (Median $= -0.2633$)

On peut lire également que la distribution est symétrique autour de 0
($Min = -Max$ et $Q_1 = -Q_3$)

Ce sont donc là des indication que notre modèle est capable d'expliquer correctement les variations de "y" en terme de "x"

Les paramètres à proprement parler

Une fois que la droite qui passe au plus près de tous les points à été trouvée, le logiciel affiche leur valeur. Ce sont ces valeurs que l'on trouve ici dans la colonne "estimate"

Coefficients:

	Estimate	Std. Error.	t	value Pr(>|t|)
(Intercept)	10.934	1.474	7.419	7.05e-07 ***
x	1.838	0.123	14.938	1.38e-11 ***

3

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

On trouve ici deux paramètres. Le premier est l'ordonnée à l'origine (intercept) C'est le "b" du $ax + b$. Le deuxième est le coefficient qui se trouve devant la variable "x" C'est la pente de la droite, le "a" du $ax + b$!

Ces valeurs sont données,

mais ATTENTION on ne commence pas la lecture de cette partie de la sortie logicielle par la colonne "estimate" !

Coefficients:

	Estimate	Std. Error.	t	value Pr(>|t|)
(Intercept)	10.934	1.474	7.419	7.05e-07 ***
x	1.838	0.123	14.938	1.38e-11 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

3

Pour interpréter correctement la valeur de ces coefficient en effet, il faut commencer par évaluer leur <u>significativité statistique</u>.

Dans le cadre d'un modèle mathématique issu d'une régression (qu'elle soit linéaire ou pas) la significativité statistique d'un paramètre mesure la probabilité, sous l'hypothèse selon laquelle sa valeur serait nulle, que la valeur soit au moins aussi grande que la valeur calculée

Autrement dit, il s'agit d'une mesure qui nous dit si la valeur calculée est significativement différente de zero.

Si tel est le cas, le paramètre en question est utile pour expliquer le y. sa p-valeur sera donc inférieure au seuil critique α Sinon, le paramètre n'est pas utile pour expliquer les données et on ne regarde alors même pas sa valeur.

Coefficients:

	Estimate	Std. Error.	t	value Pr(>|t|)
(Intercept)	10.934	1.474	7.419	7.05e-07 ***
x	1.838	0.123	14.938	1.38e-11 ***

3

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Dans notre example les deux paramètres sont statistiquement significatifs, on peut donc s'intéresser à leur valeur à proprement parler.

Il s'agit des valeurs suivantes :

Intercept = 10.934

slope (x) = 1.838

Notre <u>modèle mathématique</u> s'écrit donc :

$$\boxed{y = 1.838\,x + 10.934}$$

Ce modèle on le voit est très proche des valeurs que nous avions trouvé à la main.

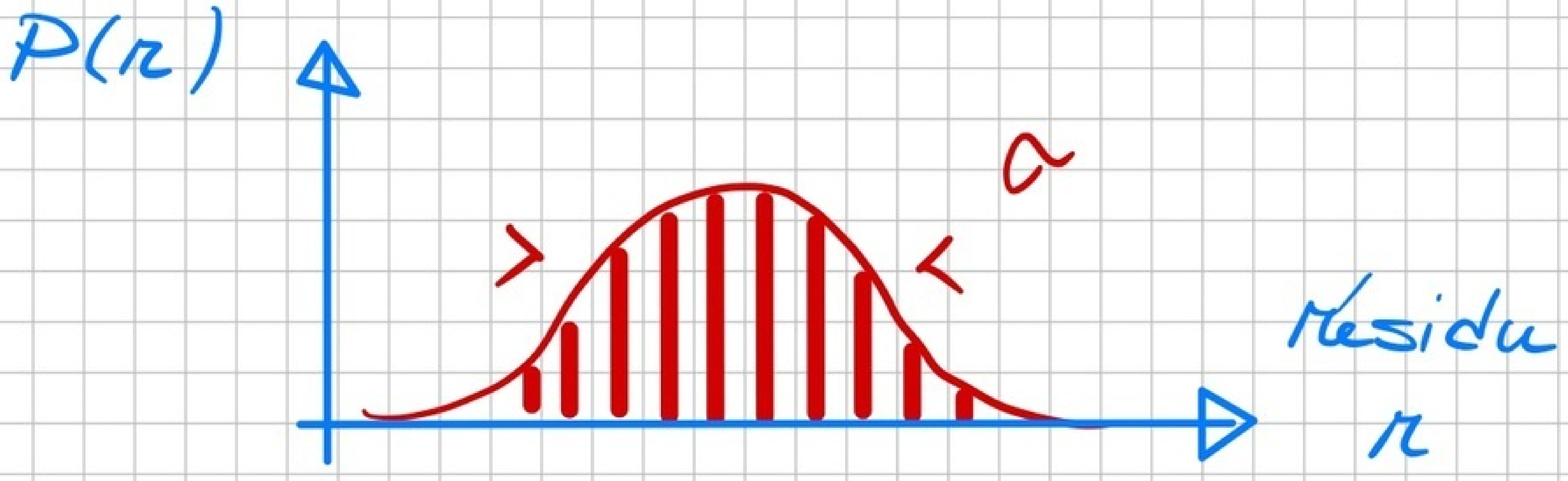
$$\boxed{y = 2x + 10}$$

La différence provient du manque de précision dont nous avons fait preuve en posant la droite à la main.

```
Residual standard error: 3.173 on 18 degrees of freedom
Multiple R-squared:  0.9254, Adjusted R-squared:  0.9212
F-statistic: 223.1 on 1 and 18 DF,  p-value: 1.382e-11
```

4

Residual Standard Error nous donne une mesure de la dispertion des valeurs des résidus autour de la moyenne



Les degrés de liberté (degree of freedom) correspondent au nombre N_T d'observation dont on a soustrait le nombre de paramètres N_{prm} du modèle

$$N_T = 20$$
$$N_{prm} = 2$$

$$20 - 2 = 18 \text{ dof}$$

Le R-carré (R-squared R^2) quant à lui est une mesure de la qualité du modèle. Il s'agit d'une valeur qui varie entre 0 et 1 :

$$0 < R^2 < 1$$

Lorsque R^2 est proche de 0, il est difficile de faire des prédictions en utilisant ce modèle. On dit alors que le pouvoir prédictif du modèle est faible.

Au contraire, lorsque R^2 est proche de 1, la qualité du modèle est bonne. Cela signifie que l'on peut s'en servir pour faire des prédictions. Autrement dit lorsque R^2 est proche de 1, le pouvoir prédictif du modèle est important.

Dans le cas des régression linéaires univariée (une seule variable prédictive) telle que celle que nous avons pris comme exemple jusqu'ici, R^2 réprésente également le rapport entre la variance expliquée par la régression et la variance totale.

La statistique F enfin (F-statistique) est elle aussi une valeur qui permet d'évaluer <u>la qualité globale du modèle</u>.

On ne l'utilise pas directement cependant comme on peut le faire avec R^2

La statistique F sert à tester l'hypothèse selon laquelle le coefficient "b" de la regression est significativement différent de 0.

$$H_0 : b = 0$$

$$H_1 : b \neq 0$$

Autrement dit on cherche a <u>savoir si la variable X est utile pour expliquer y</u>.

Dans notre exemple, la p-valeur associée à la statistique de test F est plus petite que le seuil alpha ($1.382 \cdot 10^{-11} < 0.05$) on rejette donc H_0 ce qui signifie qu'il est raisonnable de considérer le coefficient "b" comme étant différent de 0.

Autrement dit, la variable X est probablement utile pour expliquer la variable Y.

Régression linéaire multiple

Maintenant que l'on sait comment sont construit les modèles linéaires simples il devient possible de se représenter les régressions linéaires multiples.

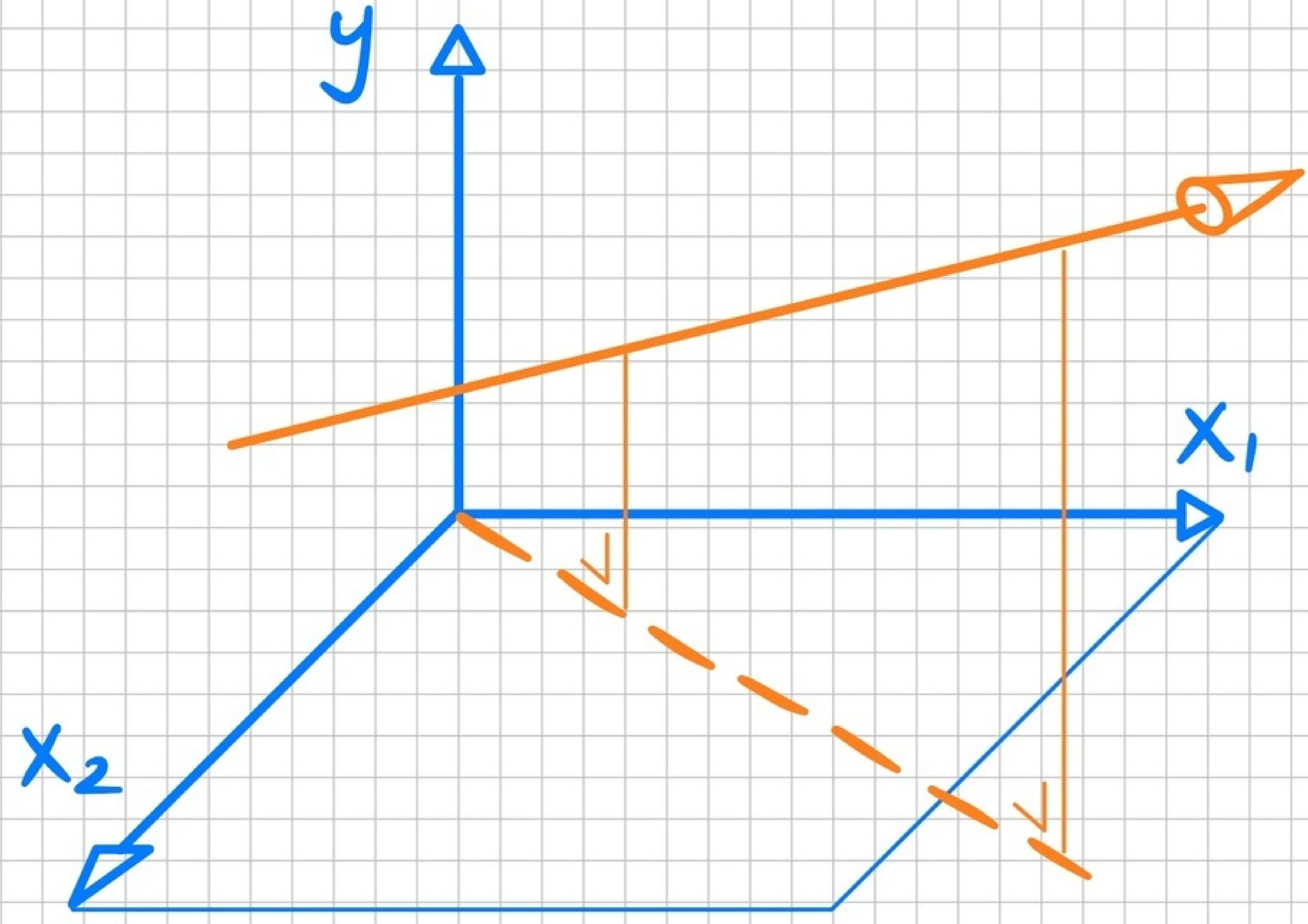
Imaginons par exemple une régression linéaire à deux variables. On note :

$$y = a_1 X_1 + a_2 X_2 + b$$

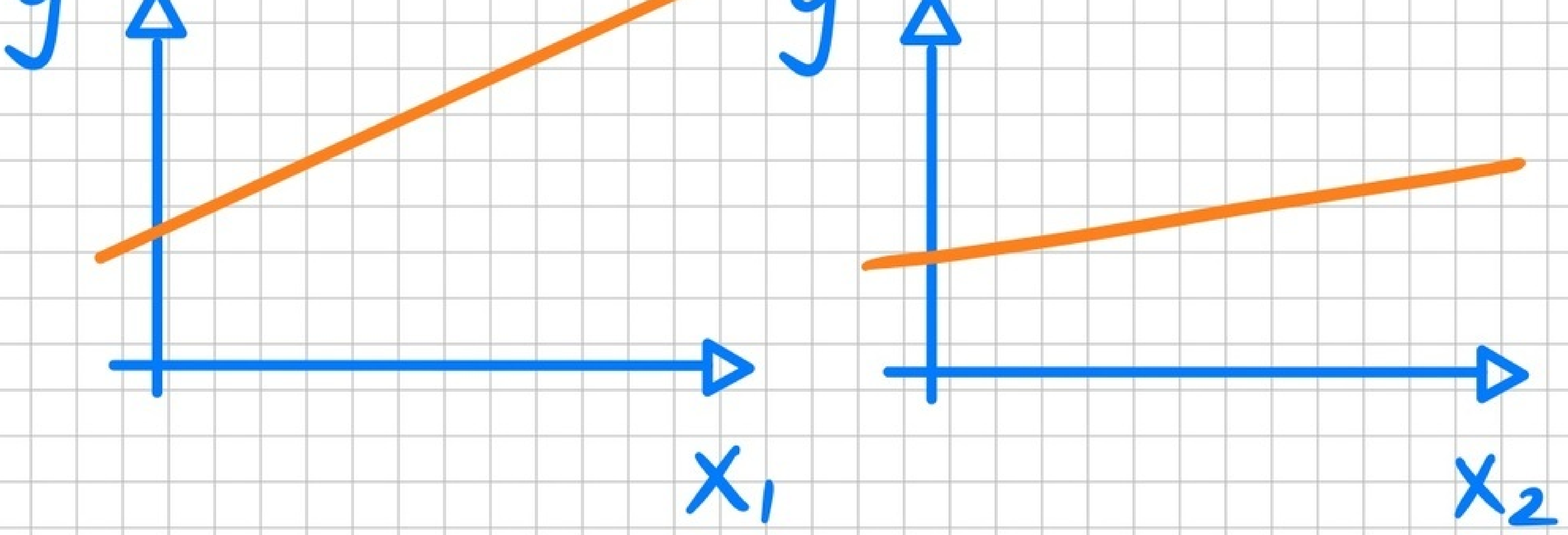
Chacune des deux variables X_1 et X_2 est ici associé à un paramètre a_1 et a_2 respectivement. Cela signifie que la variable y est proportionnelle aussi bien à X_1 que à X_2

Comme nous avons affaire ici à 2 variables X plus 1 variable y, il est possible de se représenter la situation en <u>3 dimensions</u>

Représentation graphique d'un modèle linéaire à deux variables :



Projections de la droite le long des deux axes X_1 et X_2



Du modèle à la réalité

Il existe entre le modèle et la réalité une différence importante qui ne se résume pas à une différence de valeur.

La différence fondamentale entre un modèle et la réalité réside dans la difficulté de réaliser des mesures susceptible de characteriser notre sujet d'étude.

On pense mesurer la taille des gens alors que ce que l'on mesure en réalité, c'est leur age par exemple, leur pays d'origine, leur genre ou peut être tout cela en même temps.

La réalité est infiniment complexe et variée et aucun modèle mathématique, aussi sophistiqué soit-il, ne pourra jamais expliquer la réalité dans toute sa diversité.

L'exemple qui va suivre va nous permettre d'illustrer ce point important.

Dans ce document, nous nous sommes servi d'une série de points pour calculer une droite de régression.

Nous avons d'abord réalisé une régression à la main, pour illustrer la démarche, puis nous avons utilisé un logiciel pour executer cette même démarche exactement.

Le résultat des calculs exacts réalisé par le logiciel, nous l'avons vu est la droite d'équation :

$$y = 1.838\,x + 10.934$$

Or, la droite qui a servi à produire ces point est la droite d'équation :

$$y = 2x + 10$$

Voici les instructions qui ont servi à générer la série de points dont nous nous sommes servie dans ce cours :

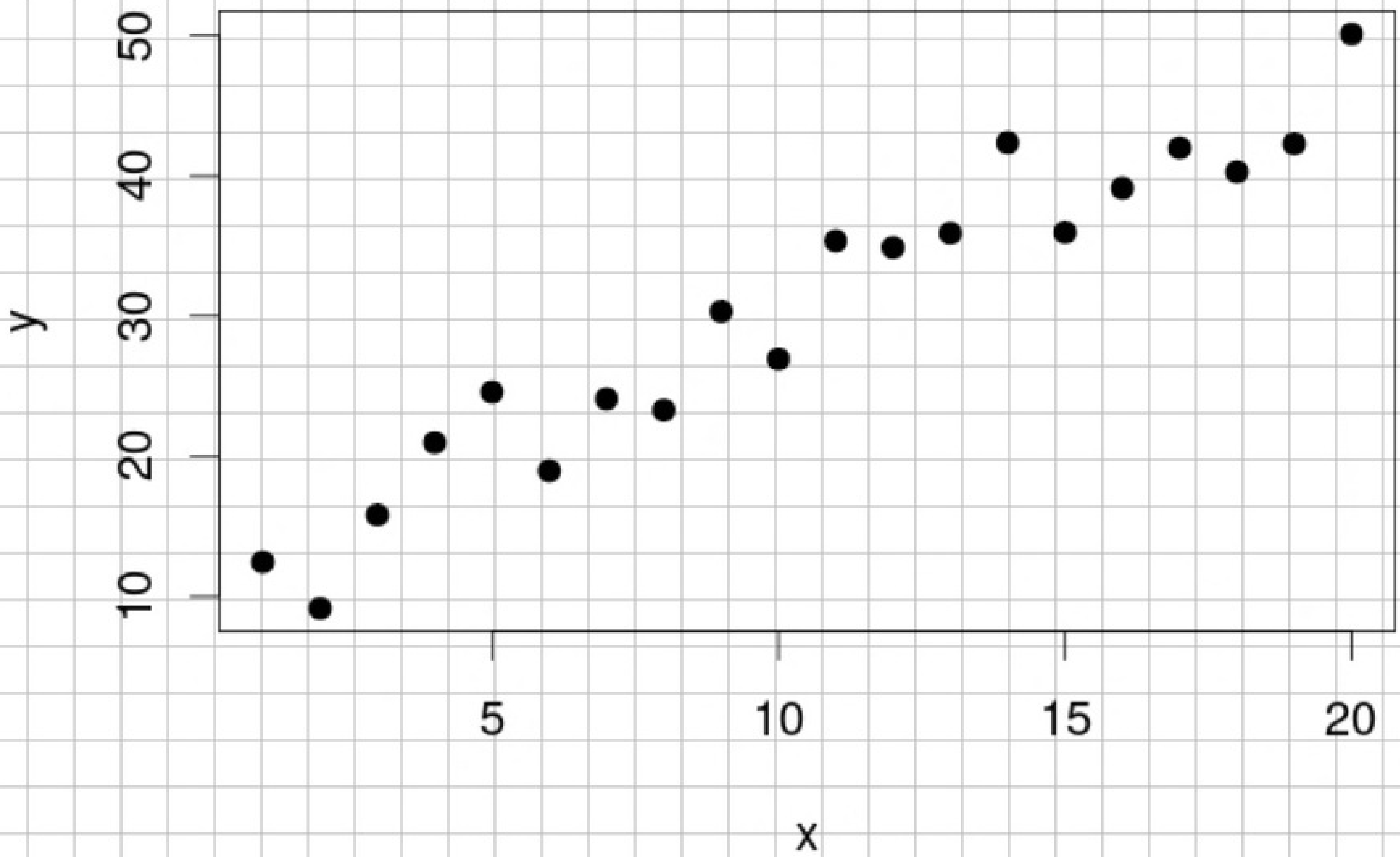
```
# On crée pour commencer
# une série de points (axe x) entre 1 et 20
x = 1:20

# Que l'on met en relation avec y
# selon une équation linéaire
# (La vrai relation entre x et y)
y = 2*x + 10

# On génère ensuite du bruit aléatoire
rnd = rnorm(20,mean=0,sd=3)

# Que l'on ajoute a y
y_rnd = y + rnd

# On affiche le résultat sur un scatterplot
plot(x,y_rnd, pch=19, col='black')
```



Les points que nous avons utilisés on été générés à l'aide d'une droite. Une certaine quantité de désordre a été ajouté ensuite

Or, le processus de régression n'a pas été capable de retrouver ces paramètre initiaux. Ils étaient comme masqué par le désordre bruit que nous avons ajouté

Si ce bruit avait été plus intense encore la droite initiale aurait totallement disparue Elle aurait été complètement couverte par le bruit ambiant. Si tel avait été le cas, l'ordinateur n'aurait pas trouvé de droite et le processus de régression aurait échoué.

Or c'est précisément ce qui se passe dans la réalité : Dans la réalité, la vraie relation entre x et y est inconnue. On essaie de la retrouver à l'aide de la méthode décrite dans ce document.

Les modèles, aussi bons soient-ils, ne sont que des représentations de la réalité. Garde bien cela présent à l'esprit à chaque fois que tu te serviras d'un modèle !

Voici mes disponibilités

Genève UNIMAIL :

Lundi et Vendredi
14h00 – 18h00

Lausanne ANTHROPOLE :

Mardi, Mercredi et Jeudi
14h00 – 18h00

Visio TELEGRAM :

Lundi – Vendredi
9h00 – 11h00
18h00 – 20h00

Réservation Calendrier
(Min 2 jours à l'avance)

Horaires WEEKEND

J'ai la possibilité de mettre à disposition des périodes durant le weekend si la situation le demande.

Mes tarifs sont alors majoré de 50%