PrivateTeacher

Cours Privés de Science

Régression Logistique

Tutoriel V 1.0

Les modèles mathématiques sont des outils dont on se sert pour se faire une représentation de la réalité.

Afin de pouvoir adapter un modèle à la réalité on utilise des valeurs que l'on peut changer librement. On appel ces valeurs <u>les paramètres du modèle</u>.

Le nombre de paramètres est différent pour chaque modèle. Les modèles linéaire par exemple, possèdent deux paramètres : la pente et l'ordonnée à l'origine.

Lorsqu'on choisi un modèle mathématique pour représenter des observations, on commence donc par ajuster les paramètres du modèle sur les observations.

Ce processus d'ajustement se nomme une régression

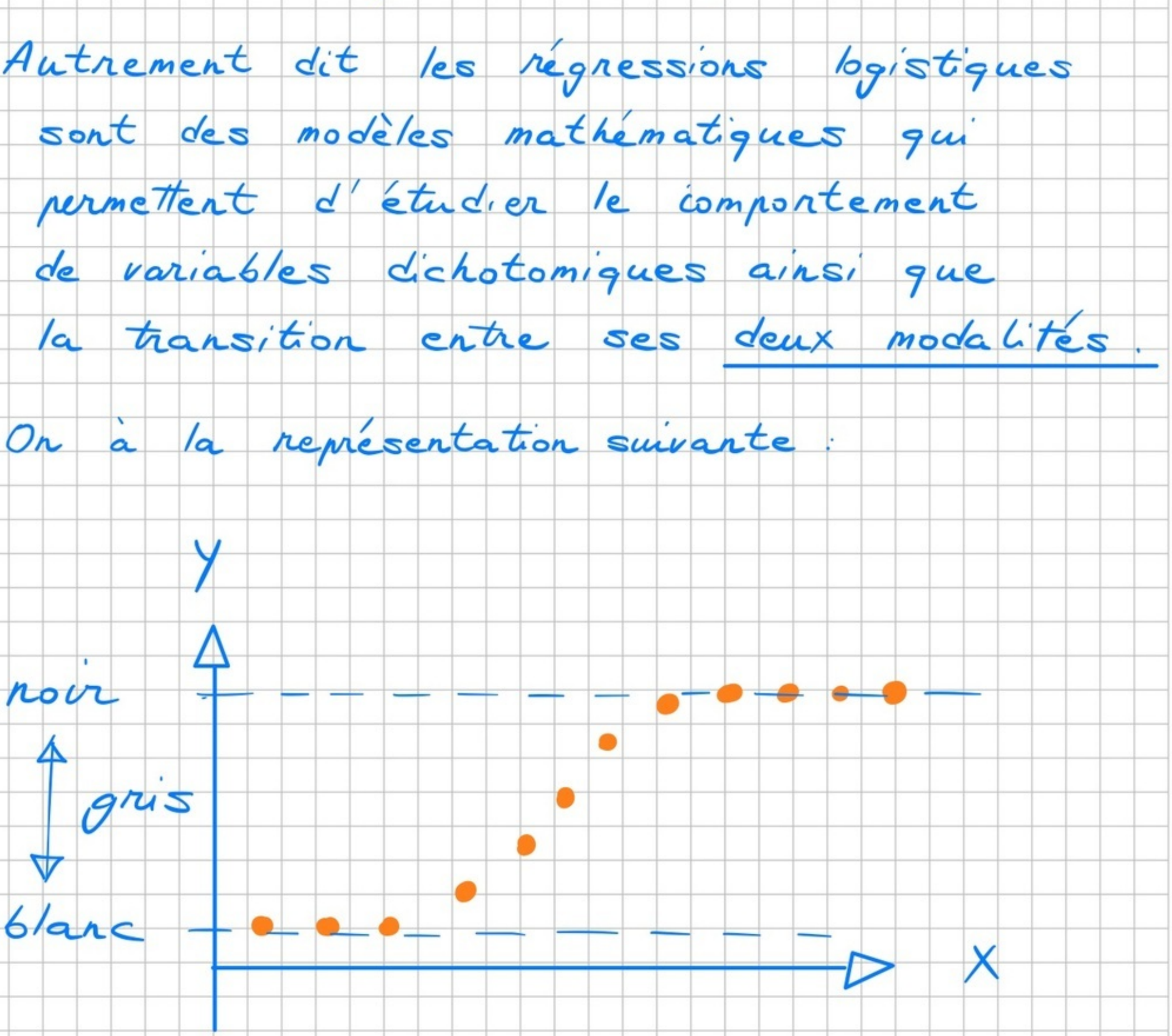
Si l'on utilise un modèle linéaire pour représenter les observations le processus d'ajustement se nomme une régression linéaire. Si l'on utilise un modèle quadratique on parlera de régression quadratique etc.

Régression logistique

On utilise les régressions logistiques lorsque l'on cherche a représenter une transition entre deux états du type juste ou faux, blanc ou noir, content - pas content

Autrement dit les régressions logistiques sont des modèles mathématiques qui permettent d'étudier le comportement de variables dichotomiques ainsi que la transition entre ses deux modalités.

On à la représentation suivante :



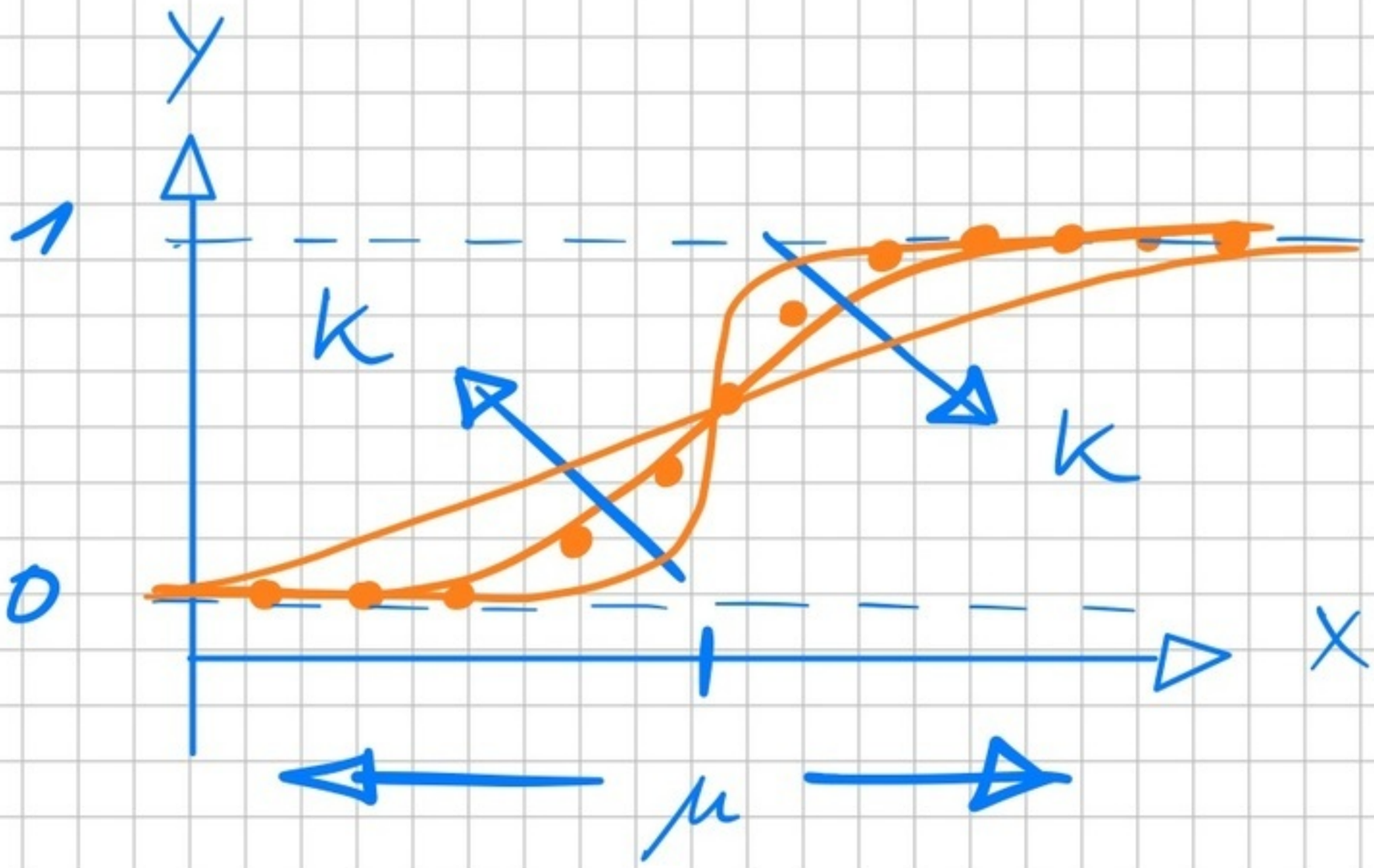
Pour représenter ce genre d'observations
on utilise une équation logistique.

Une équation logistique est une égalité
mathématique qui prend la forme suivante

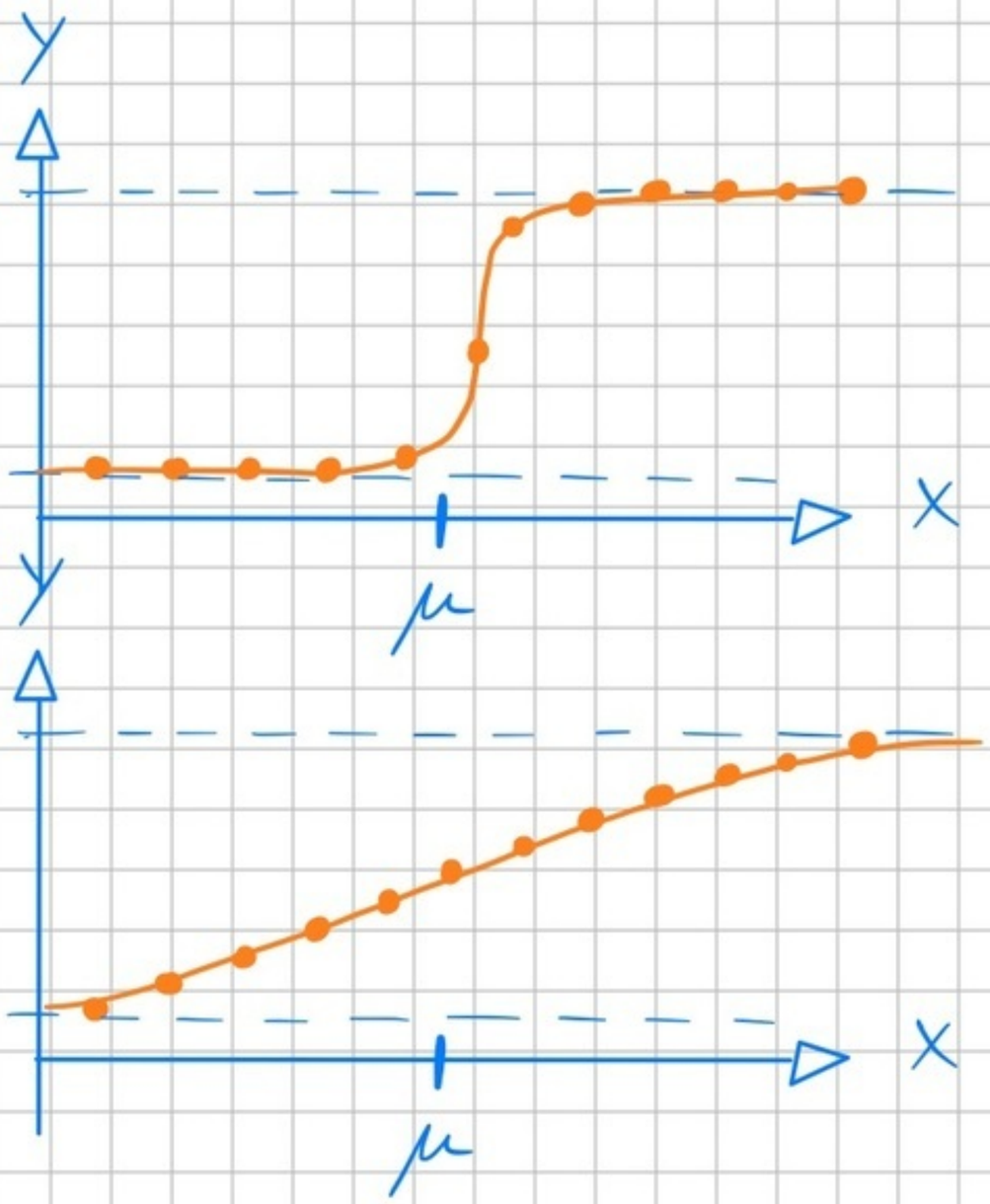
$$y = \frac{1}{1 + e^{-k(x-\mu)}}$$

Où y est la variable dont on cherche
à expliquer le comportement et où x
est la variable le long de laquelle
s'opère la transition.

Les valeurs a et b quant à elles sont
les deux paramètres de la fonction logistique
Ce sont ces deux valeurs qui donnent
leurs formes particulières aux équations
logistiques.



μ représente la valeur sur X à laquelle à lieu la transition. s quant à lui est un paramètre qui contrôle " la vitesse" à laquelle s'opère la transition :



transition abrupt
⇔ k petits

transition douce
⇔ k grand

Cette fct donne des valeurs entre 0 et 1

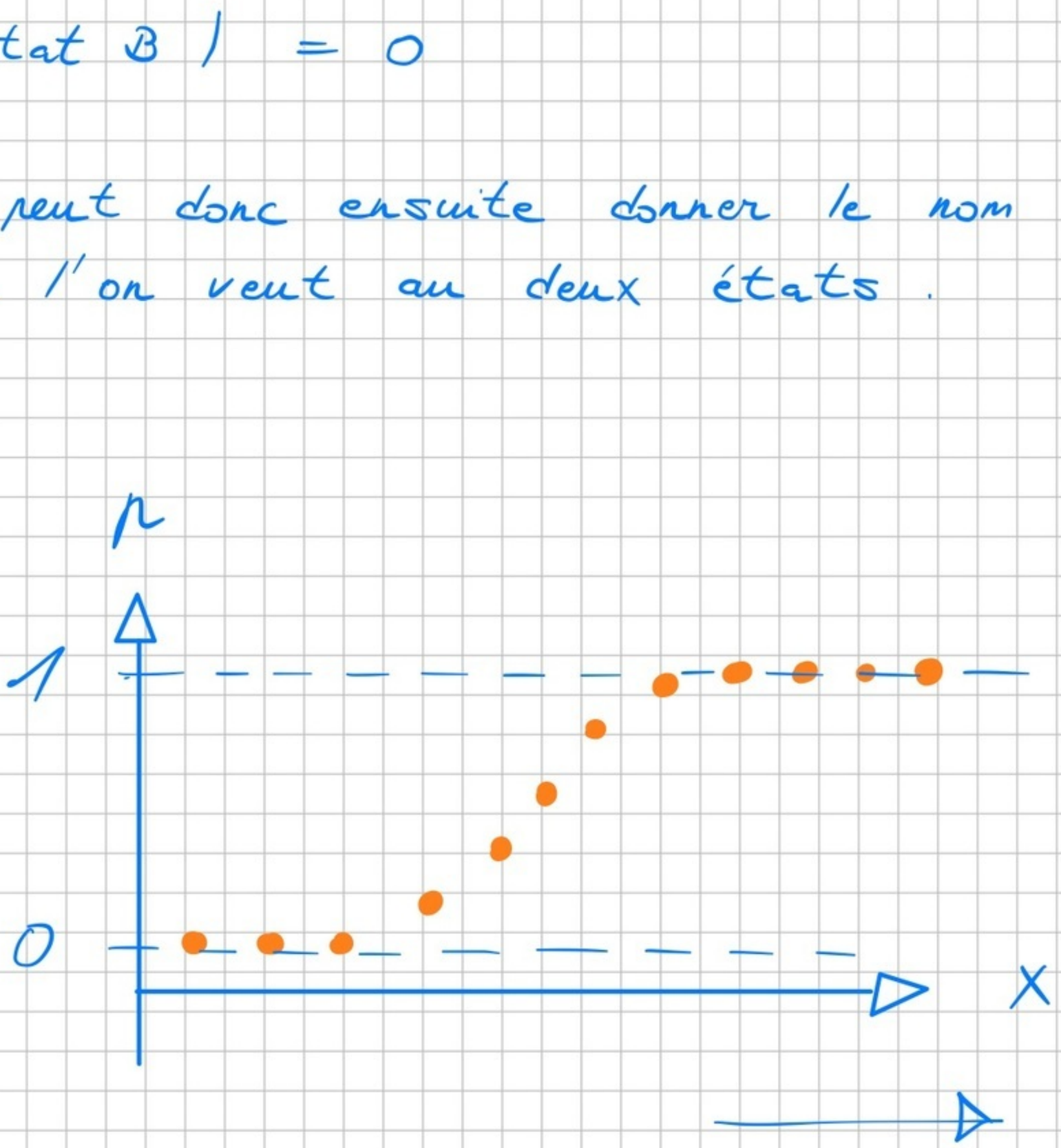
"

il s'agit donc d'une probabilité

P(état A) = 1

P(état B) = 0

On peut donc ensuite donner le nom que l'on veut au deux états.



A mesure que X augmente
la probabilité de l'événement A
devient = 1

On pourrait totallement ajuster y selon les paramètres a et b de la fonction logistique

Pour des raisons technique cependant, on préfère utiliser des modèles linéaire pour ajuster des paramètres

Cela vient du fait que l'on dispose de nombreuse méthodes pour exécuter et mener / conduire la régression, en particulier les méthodes de convergence de modèle a plusieurs paramètres sont très efficace pour les modèles linéaire.

Pour cette raison, plutôt que d'ajuster les paramètres de la fonction logistique, on va tout d'abord la linéariser.

Linéariser une fonction logistique

$$y = \frac{1}{1+e^{-k(x-\mu)}} = p$$

$$1 = p\left(1+e^{-k(x-\mu)}\right)$$

$$1 = p + pe^{-k(x-\mu)}$$

$$1-p = pe^{-k(x-\mu)}$$

$$\frac{1-p}{p} = e^{-k(x-\mu)}$$

$$\frac{p}{1-p} = e^{k(x-\mu)}$$

$$\ln\left(\frac{p}{1-p}\right) = \ln\left(e^{k(x-\mu)}\right)$$

$$\ln\left(\frac{p}{1-p}\right) = k(x-\mu)$$

$$\ln\left(\frac{p}{1-p}\right) = kx - k\mu$$

On pose : $k = a$

$-k\mu = b$

$$\ln\left(\frac{p}{1-p}\right) = ax + b$$

Autrement dit la fonction logit est linéaire en X

C'est donc de cette fonction logit dont on va se servir pour faire la regression logistic.

On a $\ln\left(\frac{p}{1-p}\right) = aX + b$

le rapport $\frac{p}{1-p}$ se nomme la cote d'un événement

La cote indique donc quelles sont les chances d'observer un événement par rapport aux chances de ne pas l'observer.

Le logarithme de la cote (la fct logit) est linéaire en X.

On peut donc faire une régression linéaire comme nos ordinateurs savent si bien les faire, puis convertir le résultat, de retour vers la fonction logistique.

On a donc les étapes suivantes

On cherche a comprendre la transition entre un état A et un état B

On utilise pour ca l'équation logistique

L'équation logistique nous permet de modéliser la transition entre deux états

Cette équation donne la probabilité y de chaque état en fonction de la valeur d'une variable X

$$y = \frac{1}{1+e^{-k(x-\mu)}}$$

Plutôt que d'ajuster y

en fonction de $\frac{1}{1+e^{-k(x-\mu)}}$

on préfère ajuster $\ln\left(\frac{y}{1-y}\right)$

en fonction de $-k(x-\mu)$

Autrement dit, on préfaire réaliser une régression linéaire sur la fonction $\ln\left(\frac{y}{1-y}\right)$ plutôt qu'une régression logistique sur y

De cette manière on peut utiliser les méthodes numériques (routine de calcul) que l'on utilise pour les régressions linéaires que l'on connait bien et qui sont déjà parfaitement au point.

On obtient donc les paramètres de la fonction logit $\pi = -kx + k\mu$

On convertit alors la fonction logit de retour vers la fonction logistique

Et c'est ainsi que l'on retrouve la probabilité qui nous intéresse en fonction de la variable X que nous avons choisi, utilisant au passage les routines de calcul numériques que l'on connait bien et qui nous permettent de faire des régressions linéaires facilement

Prenons à présent un exemple pour nous permettre de fixer les idées concrètement.

Exemple de calcul

On s'intéresse à savoir quel est le degrès de satisfaction d'un individu vis-à-vis de son salaire. On cherche également à déterminer à partir de quel montant les personnes passe de insatisfait a satisfait.

On a les données suivantes

Salaire	Satisfaction
1000	0.0000
3182	0.0000
5364	0.0022
7545	0.0126
9727	0.068
11909	0.2947
14091	0.7053
16273	0.932
18455	0.9874
20636	0.9978
22818	0.9996
25000	0.9999

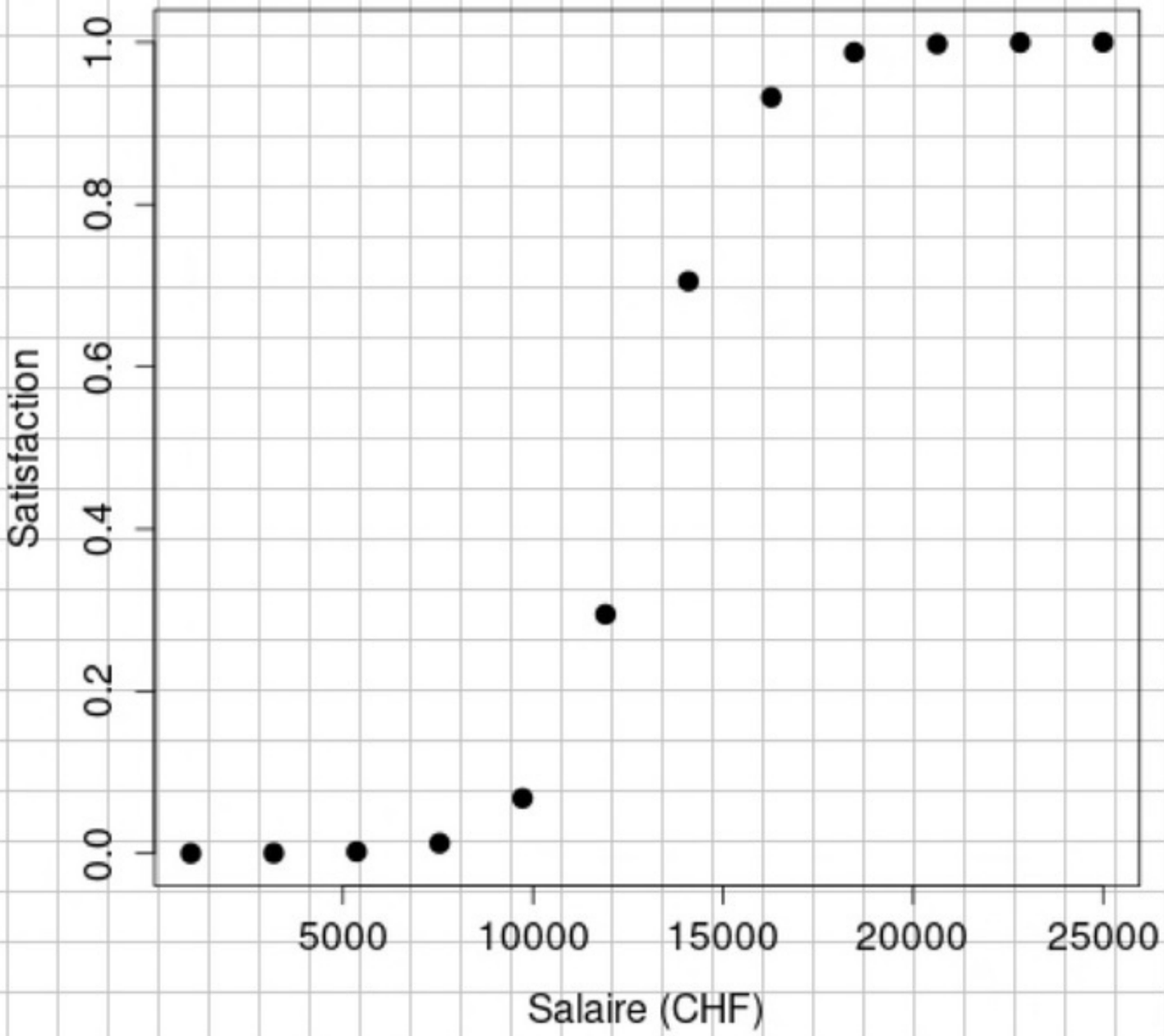
Ces données nous montrent la satisfaction des individus en fonction de leur salaire mensuel

La satisfaction de chaque individu est mesurée sur une échelle allant de 0 à 1

0 signifie insatisfait

1 signifie très satisfait

Lorsqu'on trace ces données sur un graphique on obtient la représentation suivante :



On remarque immédiatement que les petits salaires ne donnent pas beaucoup de satisfaction alors que des salaires élevés donne lieu à une grande satisfaction.

On remarque aussi qu'il existe un certain salaire au delà duquel les individus sont satisfait. Un peu comme un seuil qui une fois dépassé garanti la satisfaction de l'individu.

On s'intéresse à déterminer cette valeur de seuil précisément.

Les observations que nous avons sous les yeux ressemblent beaucoup à une fonction logistique dont l'équation, nous l'avons vu, est la suivante :

$$y = \frac{1}{1 + e^{-k(x-\mu)}}$$

Dans notre cas, la variable "x" représente le salaire des individus et la variable "y" représante leur satisfaction.

On pourrait faire une régression logistique directement sur cette fonction. Lorsque celà est possible cependant, on préfère réaliser des régressions linéaires.

C'est bien là l'utilité de linéariser la fonction logistique. Si on le fait rappelons le, on obtient la fonction logit qui elle est linéaire en X

$$\ln\left(\frac{y}{1-y}\right) = -k(x-\mu)$$

Si on applique cette transformation à nos observations, on obtient les données suivantes :

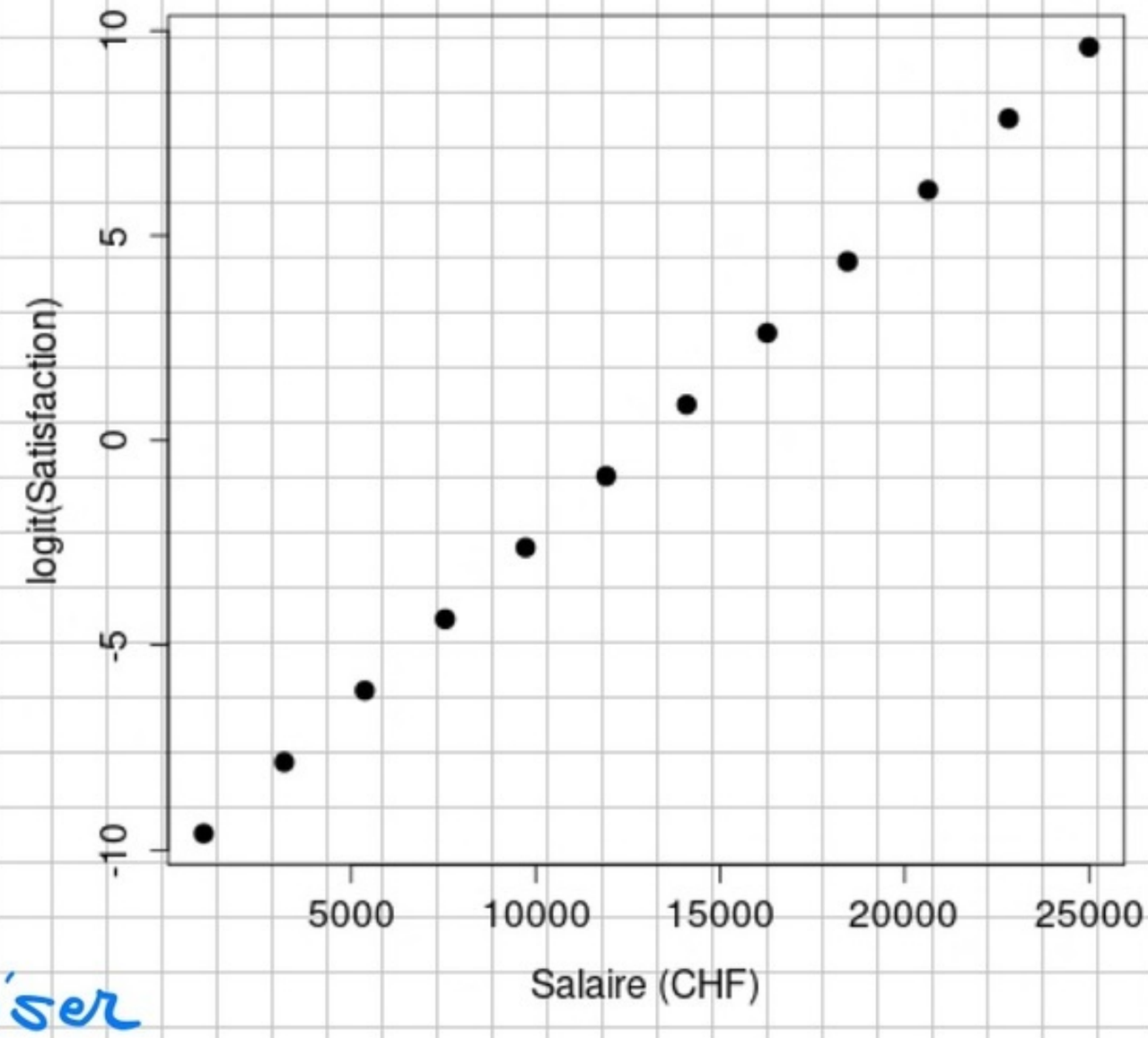
Salaire	Logit(Satis)
1000	-9.6
3182	-7.8545
5364	-6.1091
7545	-4.3636
9727	-2.6182
11909	-0.8727
14091	0.8727
16273	2.6182
18455	4.3636
20636	6.1091
22818	7.8545
25000	9.6

Les salaires restent les mêmes, on a simplement transformé les valeurs de l'échelle des satisfactions :

y devient $\ln\left(\frac{y}{1-y}\right)$

$= logit(y)$

Lorsqu'on applique cette transformation on obtient le graphique suivant :



On vient tout juste de linéariser nos observations.

On peut dès lors réaliser facilement une régression linéaire à l'aide des outils (méthodes numériques) que l'on connait bien. C'est d'ailleurs là tout l'intéret de linéariser nos observations

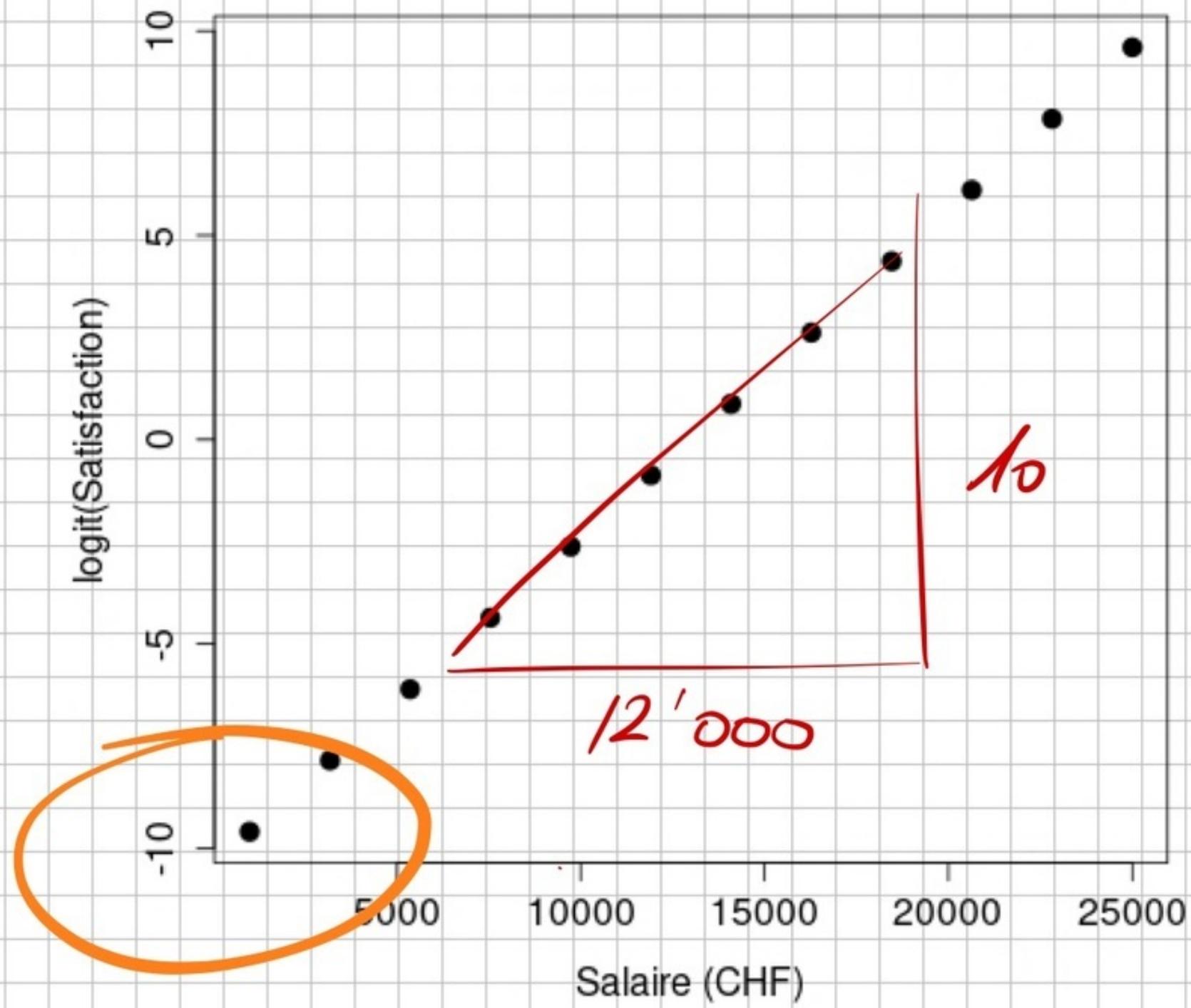
L'intérêt ne semble pas évident lorsqu'il n'y a qu'une seule variable X mais devient important lorsque l'on cherche a ajuster une fonction logistique sur plusieurs variables comme nous allons le voir dans un instant.

Reprenons notre série de points qui s'aligne sur une droite et laissons l'ordinateur (R) trouver son equation.

Les valeurs exacte ne sont pas très importantes car c'est surtout la démarche qui nous intéresse ici.

On peut donc parfaitement trouver les paramètres de la droite à la main

Si l'on fait cela on trouve les paramètres suivants :



Ordonnée à l'origine $= -10$

$$\text{Pente} = \frac{10}{12\,000} = 0.0008$$

On a donc l'équation suivante :

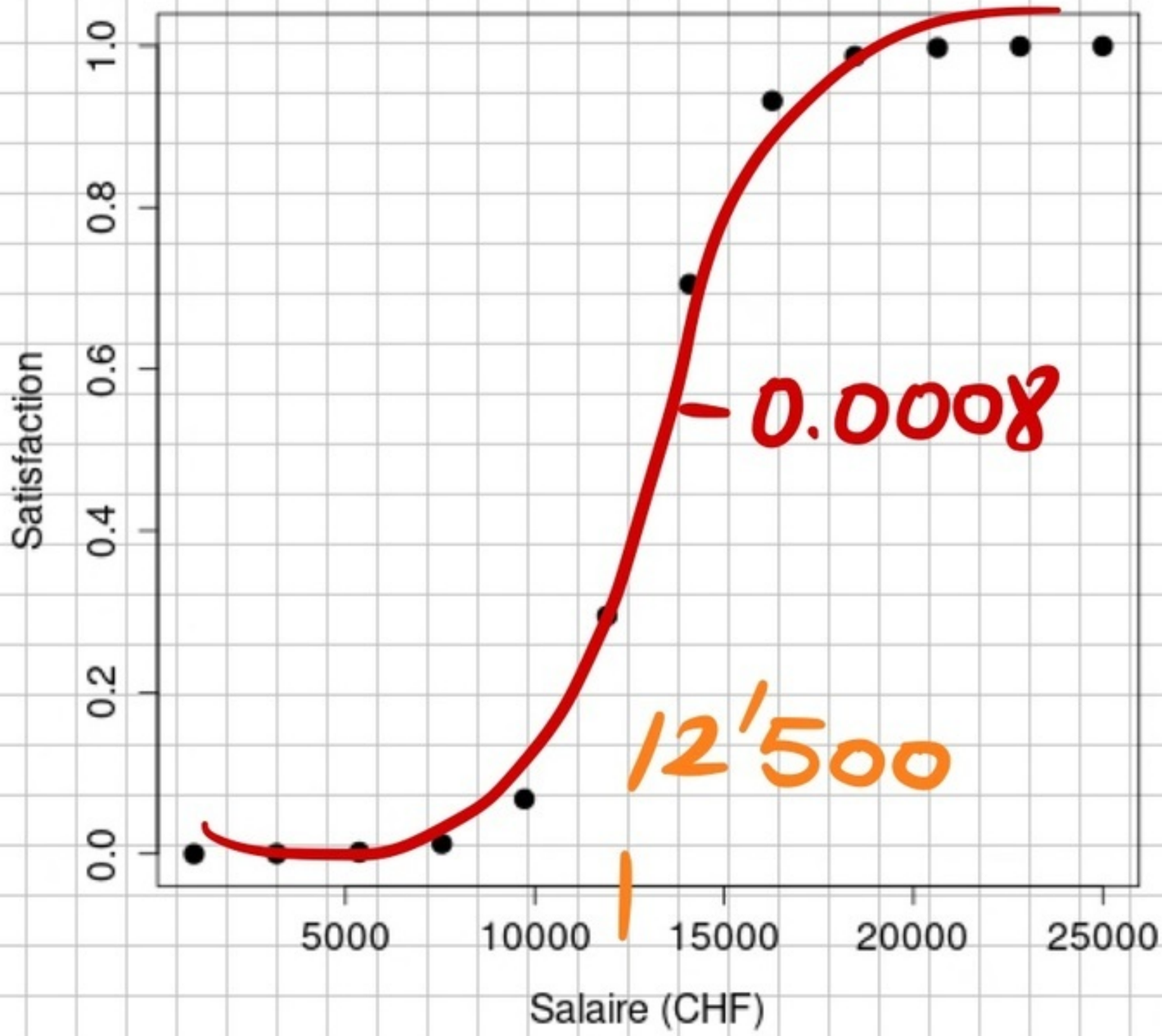
$$\log\left(\frac{y}{1-y}\right) = 0.0008\,X - 10$$

On peut dès lors retrouver notre fonction de départ

$$y = \frac{1}{1 + e^{-k(x-\mu)}}$$

$$y = \frac{1}{1 + e^{0.0008x - 10}}$$

$$y = \frac{1}{1 + e^{-0.0008(x + 12500)}}$$



Autrement dit le salaire limite au delà duquel les individus sont satisfait est de 12'500 CHF.

Voyons à présent les résultats d'une régression logistique faite sur un ordinateur.

```
Call:
glm(formula = Satis ~ Salaire, family = "binomial", data =
data)

Deviance Residuals:
      Min         1Q     Median         3Q        Max
-0.0036100  -0.0003127   0.0000000   0.0003127   0.0036100

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.039e+01  7.686e+00  -1.352    0.176
Salaire      7.995e-04  5.824e-04   1.373    0.170

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1.2865e+01  on 11  degrees of freedom
Residual deviance: 2.7407e-05  on 10  degrees of freedom
AIC: 5.7404

Number of Fisher Scoring iterations: 8
```

On retrouve là les paramètres que nous connaissons

$b = \text{intercept} = -1.039 \cdot 10^{1} \simeq \underline{-10}$

$a = \text{slope} = 7.995 \cdot 10^{-4} \simeq \underline{0.0008}$

Il s'agit bien des paramètres de la droite de la fonction $\text{logit}(y)$

$$\text{logit}(y) = \log\left(\frac{y}{1-y}\right) = aX + b$$

$$= 0.0008X - 10$$

Ils sont accompagné cette fois de leur terme d'erreur et de leur significativité statistique.

Cette fonction logit nous permet quant à elle de formuler le modèle logistique qui nous intéresse depuis le départ

$$\log\left(\frac{y}{1-y}\right) = aX + b$$

$$y = \frac{1}{1+e^{-(aX+b)}}$$

$$= \frac{1}{1+e^{-a(X+b/a)}}$$

$$= \frac{1}{1+e^{-0.0008\left(X+\frac{10}{0.0008}\right)}}$$

$$= \frac{1}{1+e^{-0.0008(X+12\,500)}}$$

Salaire	Satisfaction
1000	0.0000
3182	0.0000
5364	0.0022
7545	0.0126
9727	0.068
11909	0.2947
14091	0.7053
16273	0.932
18455	0.9874
20636	0.9978
22818	0.9996
25000	0.9999