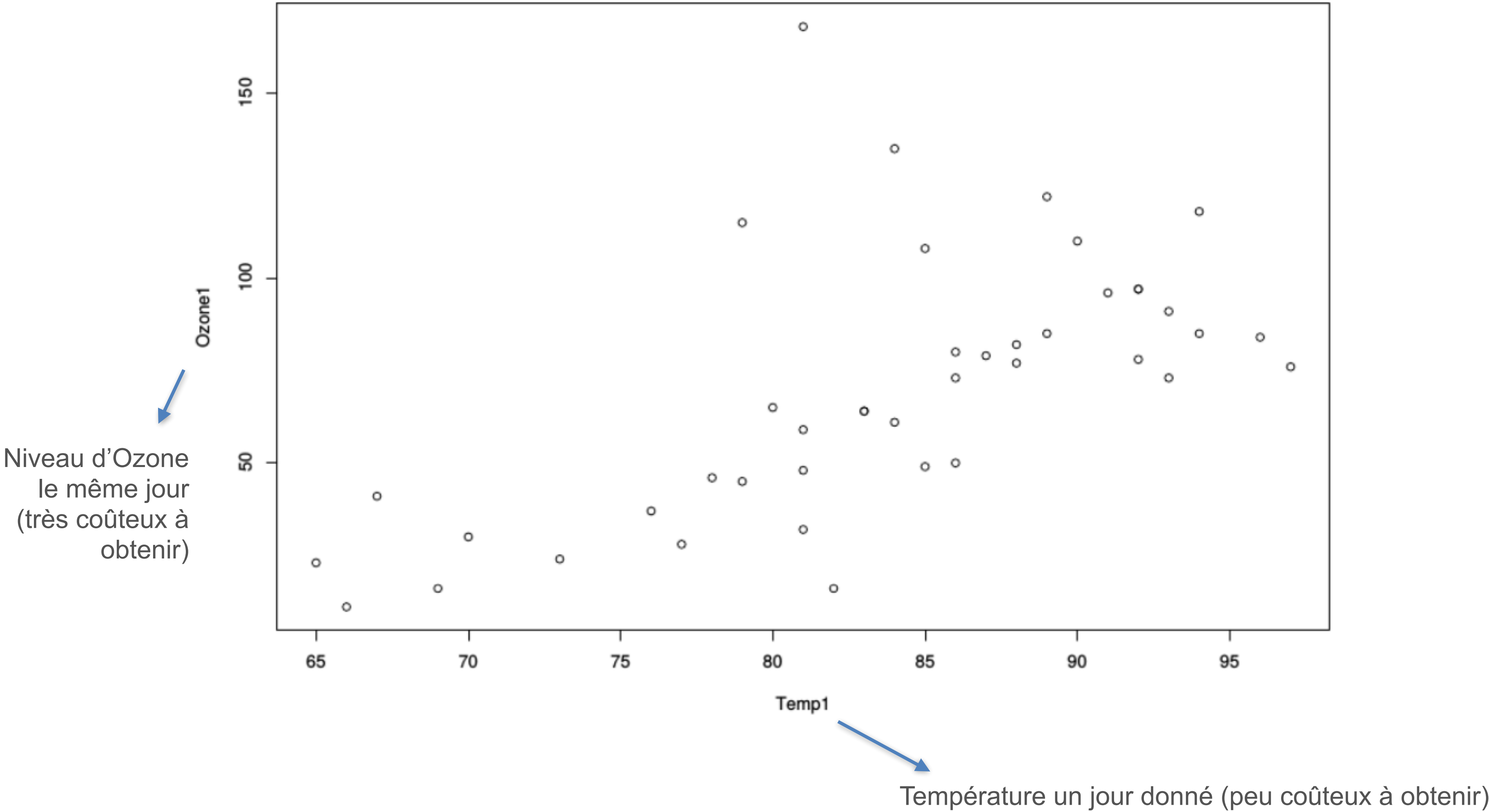


# Fondements statistiques de l'apprentissage automatique

## Chapitre 2 : Régression linéaire

On s'intéresse à l'étude de phénomènes supposés linéaires, par exemple :



**FS-AA - chapitre 2 : Régression linéaire → 2.1 Régression linéaire 1D**

On note  $Y$  la variable aléatoire réelle à expliquer (ou encore de réponse, dépendante) et  $X$  la variable explicative (ou encore déterministe, de contrôle) ou effet fixe ou facteur contrôlé. Le modèle revient à supposer, qu'en moyenne, l'estimation  $\mathbb{E}(Y)$ , est une fonction affine de  $X$ .

$$\mathbb{E}(Y) = f(X) = \beta_0 + \beta_1 X.$$

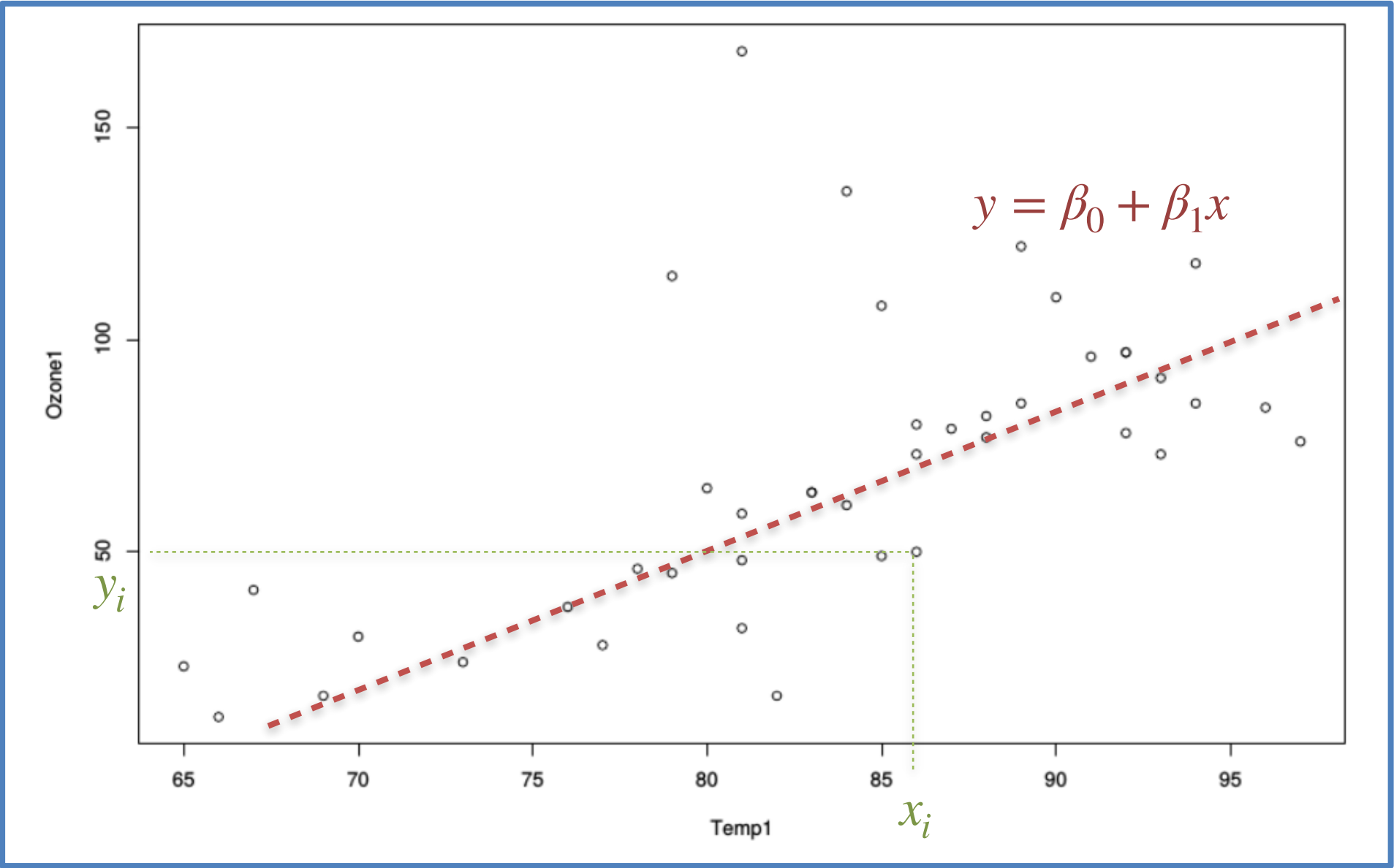
Pour une séquence d'observations aléatoires identiquement distribuées  $\{(y_i, x_i), i = 1, \dots, n\}$ , avec  $n > 2$  et les  $x_i$  non tous égaux, le modèle s'écrit à partir des observations :

$$y_i = \beta_0 + \beta_1 x_i + u_i, i = 1, \dots, n$$

ou bien sous forme matricielle :

$$\begin{aligned} \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} &= \begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + \begin{bmatrix} u_1 \\ \vdots \\ u_n \end{bmatrix}, \\ \mathbf{y} &= \mathbf{X}\boldsymbol{\beta} + \mathbf{u} \end{aligned}$$

où le vecteur  $\mathbf{u}$  contient les erreurs.

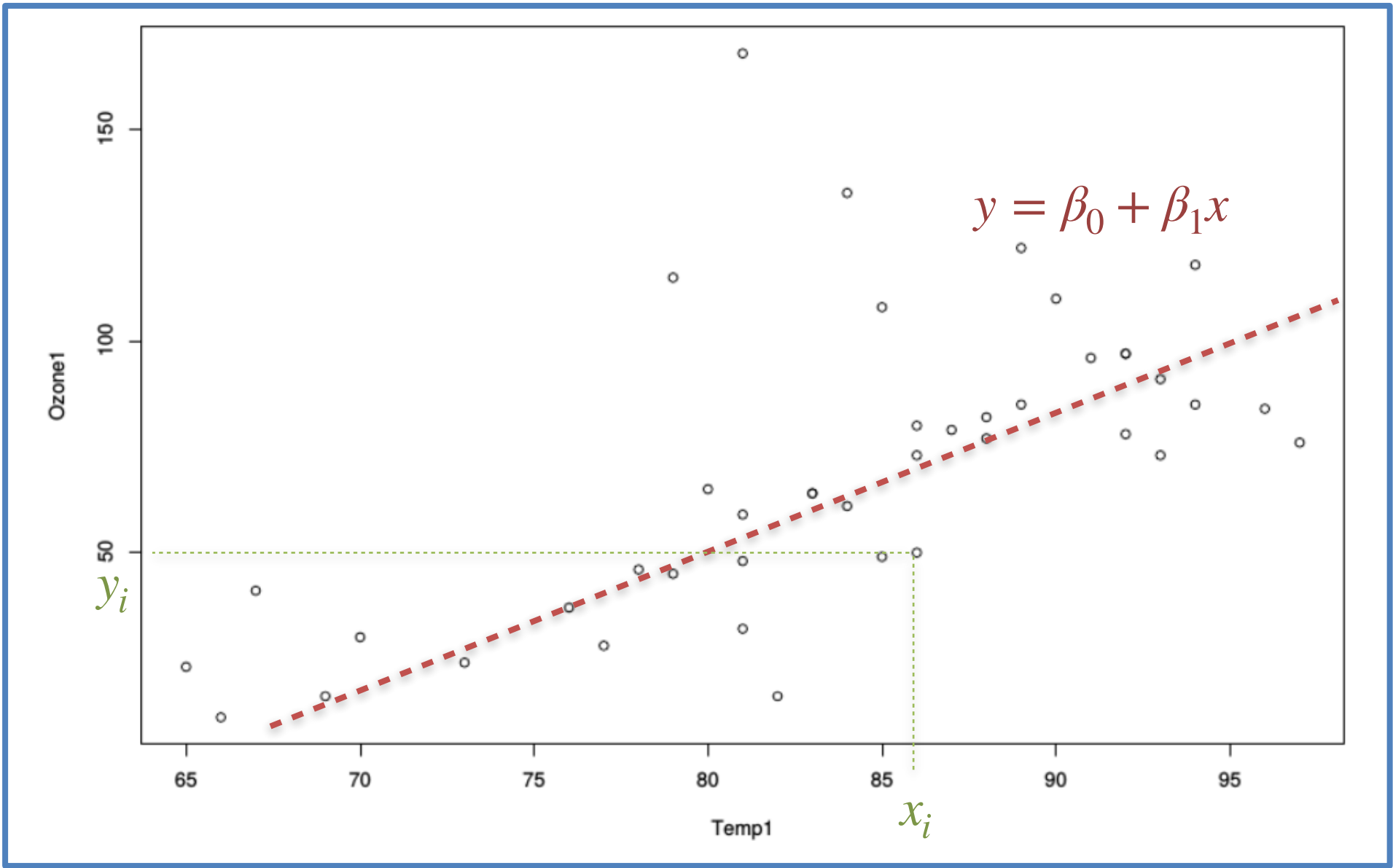


$$\begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + \begin{bmatrix} u_1 \\ \vdots \\ u_n \end{bmatrix},$$
$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}$$

- Les hypothèses relatives à ce modèle sont les suivantes :
- la distribution de l’erreur  $\mathbf{u}$  est indépendante de  $X$  ou bien  $X$  est fixe.
  - l’erreur est centrée et de variance constante (homoscédasticité) :

$$\forall i = 1, \dots, n : E(u_i) = 0, Var(u_i) = \sigma_u^2.$$

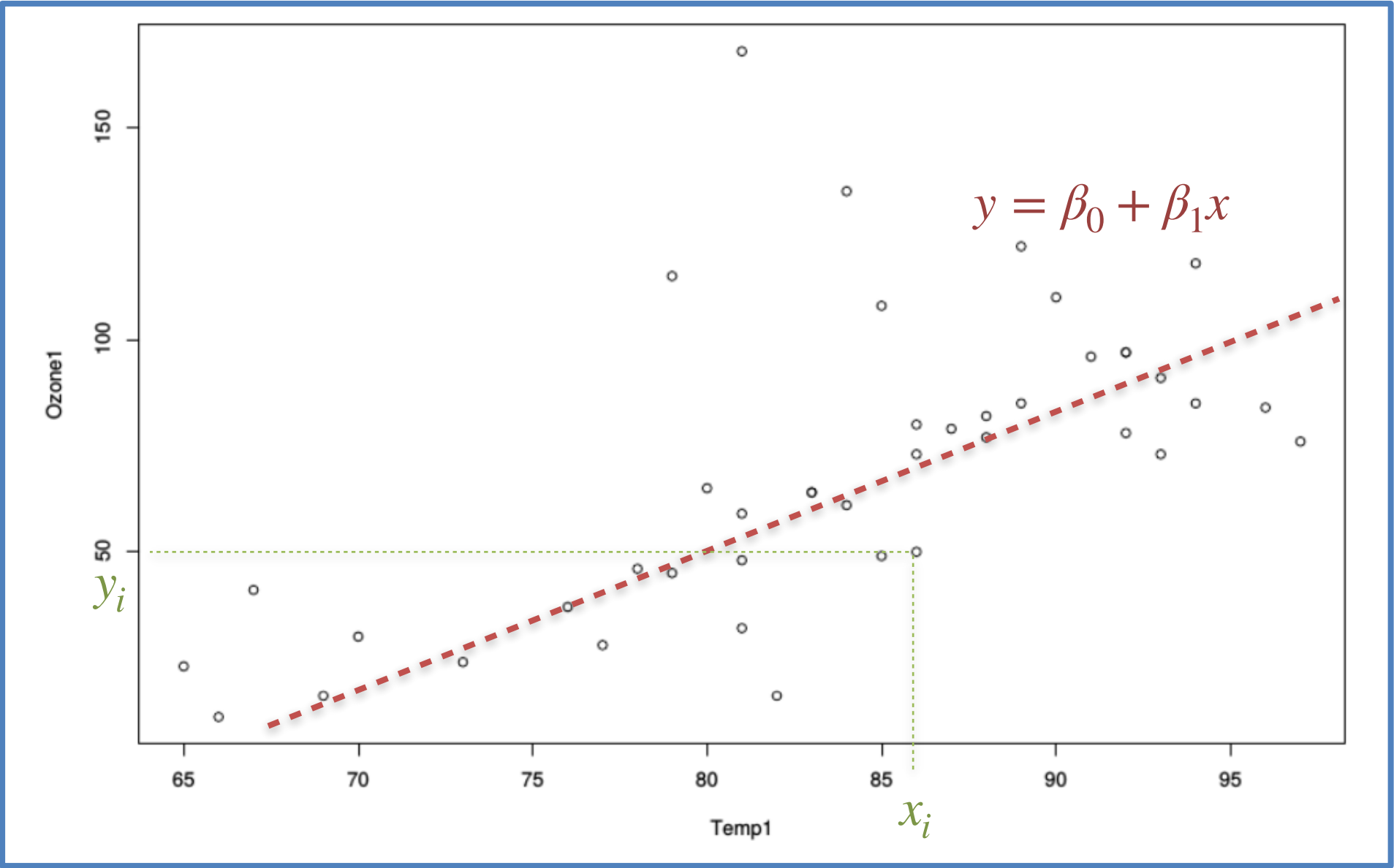
- $\beta_0$  et  $\beta_1$  sont constants, il n’y a pas de rupture du modèle.
- Hypothèse complémentaire pour les inférences :  $u \sim \mathcal{N}(0, \sigma_u^2 \mathbf{I}_p)$ .



**FS-AA - chapitre 2 : Régression linéaire → 2.1 Régression linéaire 1D**

L'estimation des paramètres  $\beta_0, \beta_1, \sigma_u^2$  peut être obtenue en minimisant la somme des carrés des écarts entre observations et modèle (moindres carrés). Pour un jeu de données  $\{(y_i, x_i), i = 1, \dots, n\}$ , le critère des moindres carrés s'écrit :

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



**FS-AA - chapitre 2 : Régression linéaire → 2.1 Régression linéaire 1D**

L'estimation des paramètres  $\beta_0, \beta_1, \sigma_u^2$  peut être obtenue en minimisant la somme des carrés des écarts entre observations et modèle (moindres carrés). Pour un jeu de données  $\{(y_i, x_i), i = 1, \dots, n\}$ , le critère des moindres carrés s'écrit :

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

Pour minimiser ce critère, on pose :

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$$

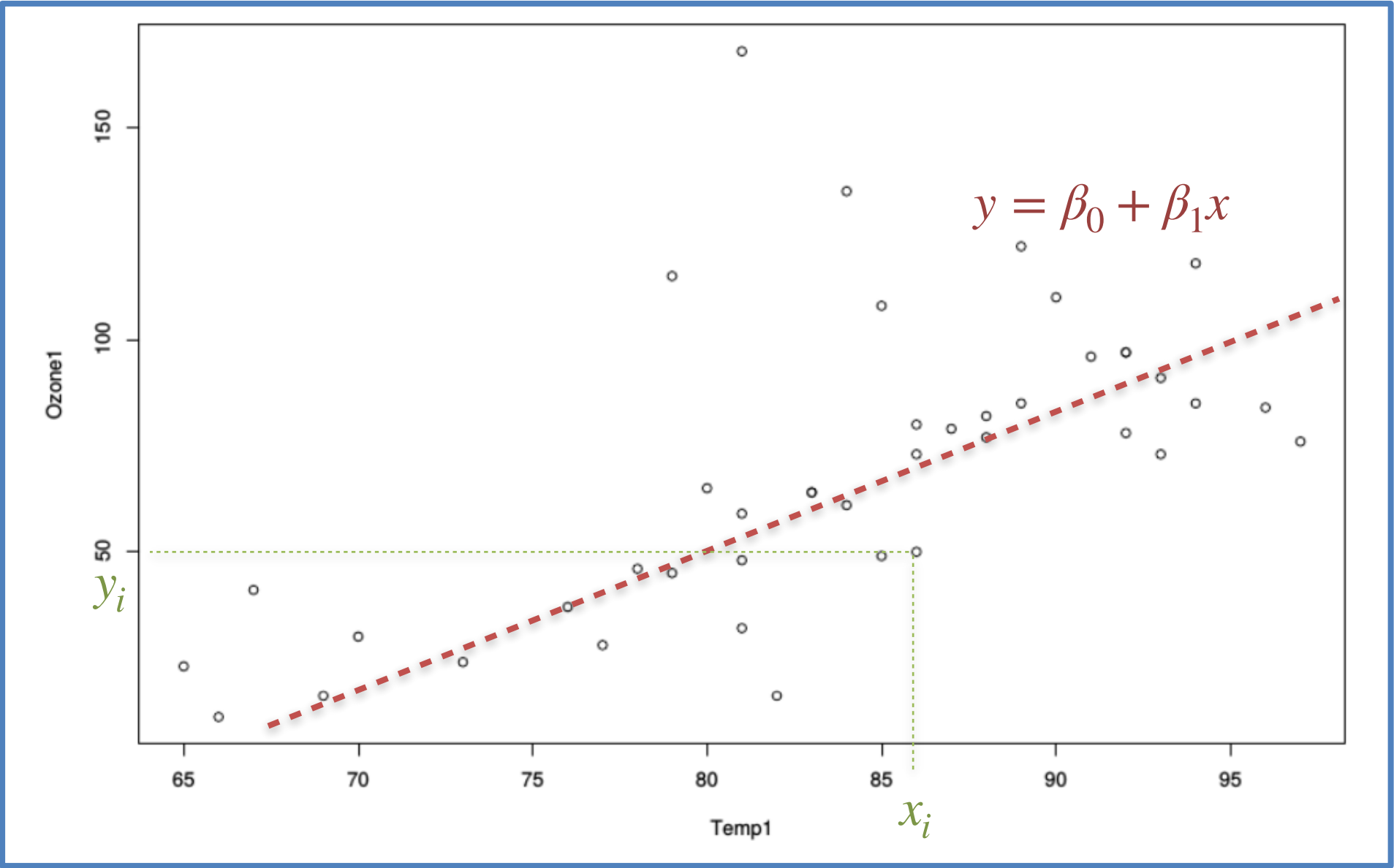
$$\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i$$

$$s_x^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$$

$$s_y^2 = \frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2$$

$$s_{xy} = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$$

$$r = \frac{s_{xy}}{s_x s_y}$$





L'estimation des paramètres  $\beta_0, \beta_1, \sigma_u^2$  peut être obtenue en minimisant la somme des carrés des écarts entre observations et modèle (moindres carrés). Pour un jeu de données  $\{(y_i, x_i), i = 1, \dots, n\}$ , le critère des moindres carrés s'écrit :

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

Pour minimiser ce critère, on pose :

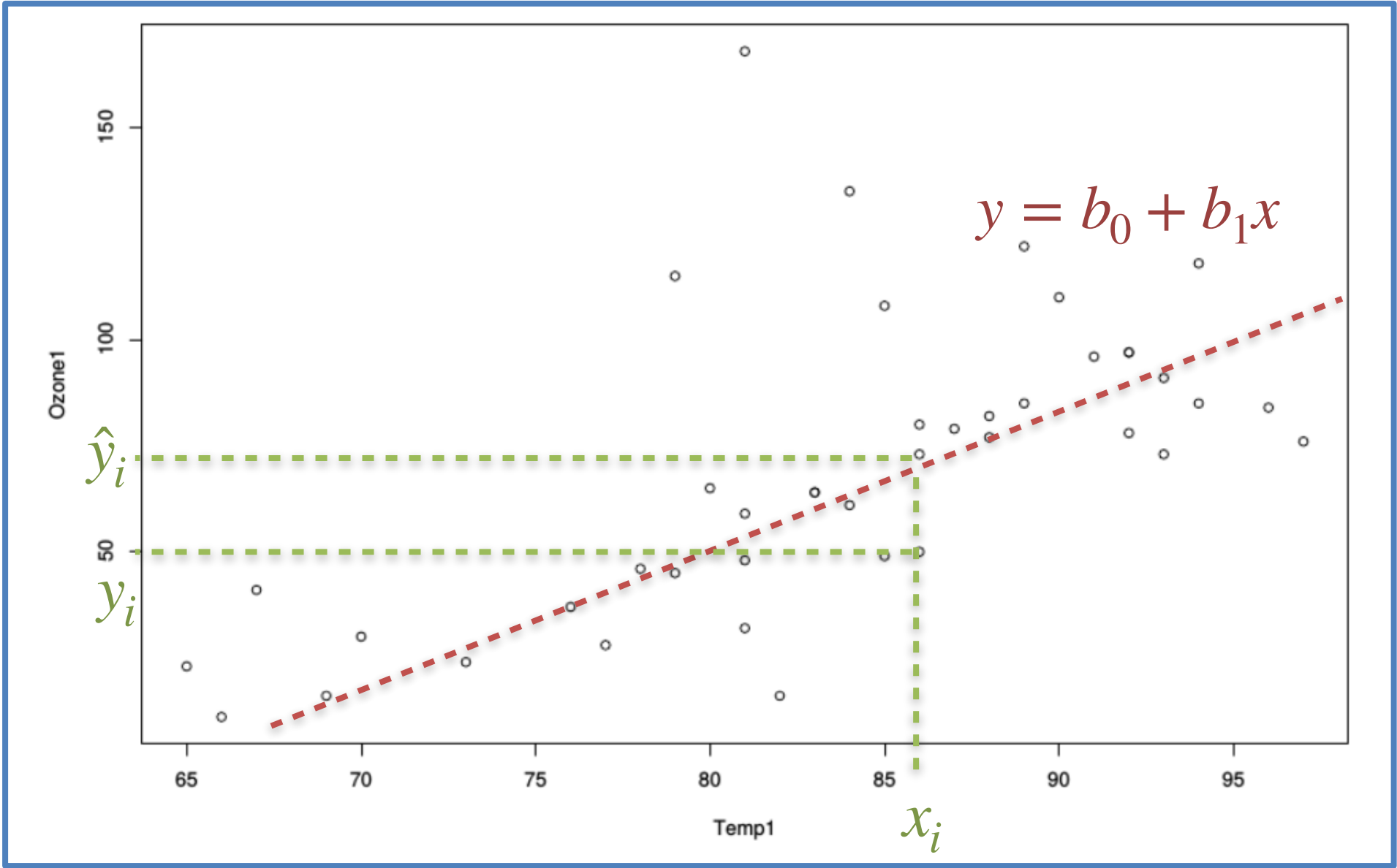
$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$$
$$\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i$$
$$s_x^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$$

$$s_y^2 = \frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2$$
$$s_{xy} = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$$
$$r = \frac{s_{xy}}{s_x s_y}$$

On peut alors montrer que les estimateurs de  $\beta_0$  et  $\beta_1$  au sens des moindres carrés sont :

$$b_1 = \frac{s_{xy}}{s_x^2},$$
$$b_0 = \bar{y} - b_1 \bar{x}.$$

bo et b1 : estimateurs de la réalité (estimation car tjrs des erreurs ex aléas)



À chaque valeur  $x_i$  de  $X$  correspond la valeur estimée (ou prédite, ajustée) de  $Y$  :  $\hat{y}_i = b_0 + b_1 x_i$  avec  $e_i = y_i - \hat{y}_i$

Niveau d’incertitude lié à l’estimation de  $b_0$  et  $b_1$

une hypothese a été faite sur les résidus  $e \sim \mathcal{N}(0, \sigma_u^2 \mathbf{I}_p)$

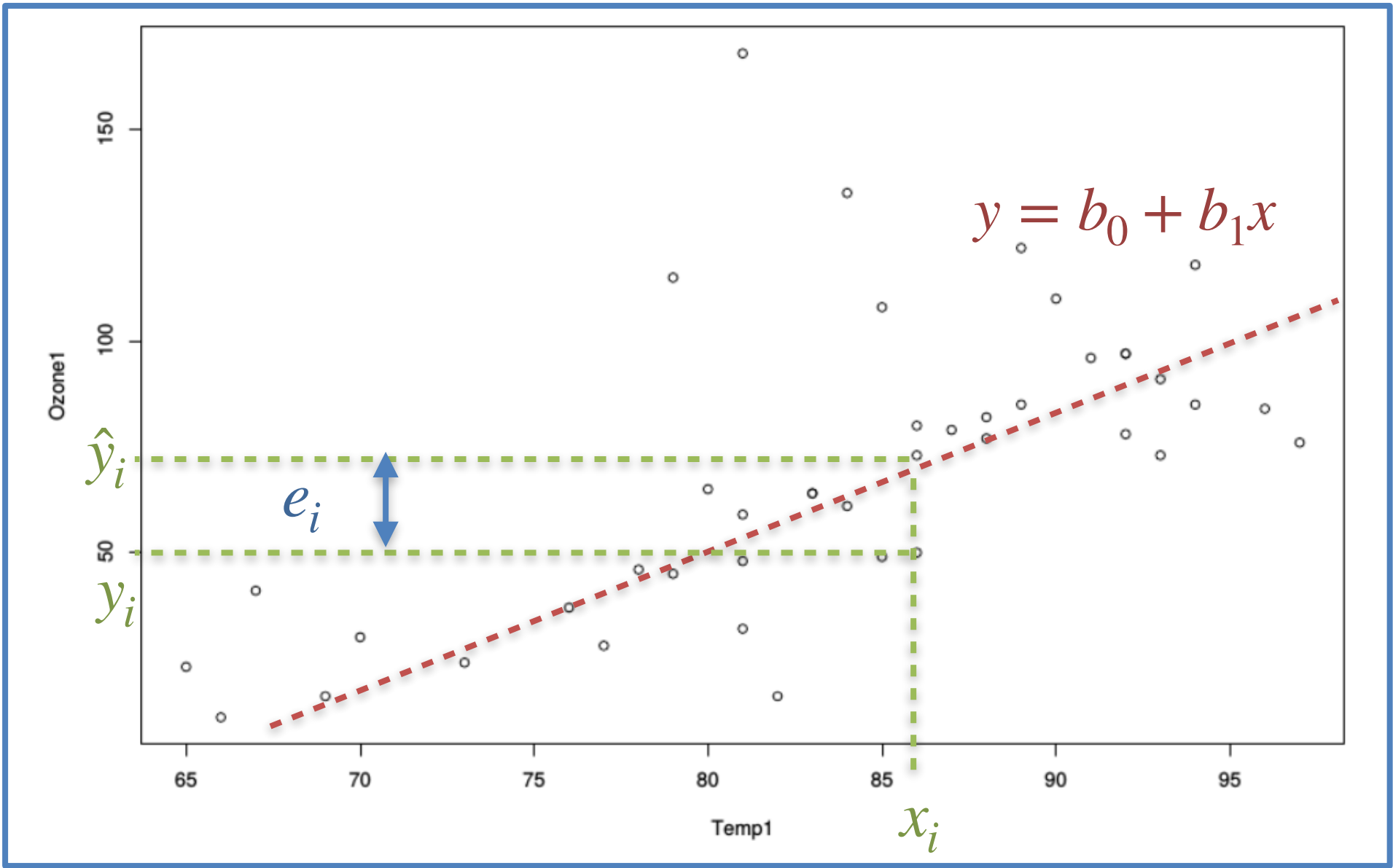
On peut alors montrer que :

$$(b_0 - \beta_0) \Big/ s \left( \frac{1}{n} + \frac{\bar{x}^2}{(n-1)s_x^2} \right)^{1/2}$$

et

$$(b_1 - \beta_1) \Big/ s \left( \frac{1}{(n-1)s_x^2} \right)^{1/2}$$

suivent des lois de Student à  $(n - 2)$  degrés de liberté. Ceci permet de tester l’hypothèse de nullité d’un de ces paramètres à partir de tests d’hypothèses. On va par exemple tester si le  $b_1$  obtenu est significativement différent de 0, en fonction d’un coefficient  $\alpha$  qui représente la probabilité avec laquelle on accepte de se tromper. Typiquement  $\alpha$  correspond à 5% de chances de se tromper, ci qui est raisonnablement faible (voir le cours de Statistique pour aller plus loin). Notons, que si  $b_1$  est significativement différent de 0, on peut considérer qu’il existe une relation de dépendance entre les  $x_i$  et les  $y_i$ .



les estimateurs de  $\beta_0$  et  $\beta_1$  au sens des moindres sont :

$$\begin{aligned} b_1 &= \frac{s_{xy}}{s_x^2}, \\ b_0 &= \bar{y} - b_1 \bar{x}. \end{aligned}$$



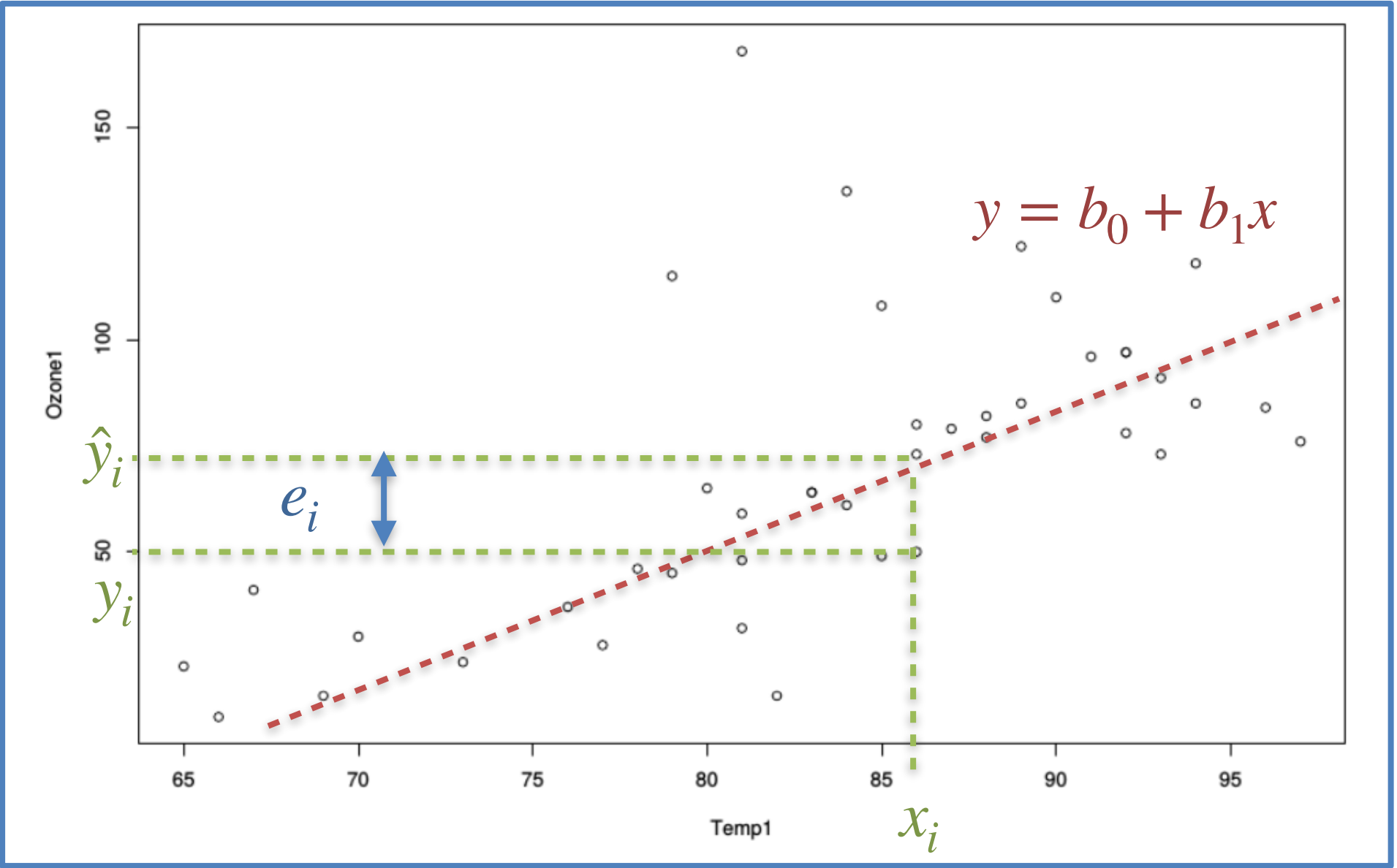
Niveau d’incertitude lié à l’estimation de  $b_0$  et  $b_1$

Il est de même possible de construire des intervalles de confiance pour les valeurs de  $b_0$  et  $b_1$ , toujours en fonction d’un niveau de confiance dépendant de  $\alpha$  :

$$b_0 \pm s \left( \frac{1}{n} + \frac{\bar{x}^2}{(n-1)s_x^2} \right)^{1/2} t_{n-2}(\alpha/2)$$
$$b_1 \pm s \left( \frac{1}{(n-1)s_x^2} \right)^{1/2} t_{n-2}(\alpha/2)$$

où  $t_\nu(\alpha)$  est la distribution de Student à  $\nu$  degrés de liberté (voir appendice A). En observant bien ces intervalles de confiance ainsi que les distributions de Student, il est intéressant de noter que plus on a d’observations  $n$ , plus les intervalles de confiances sont resserés autour des  $b_0$  et  $b_1$  estimés. Plus on dispose d’information, moins le risque d’erreur est en effet grand par rapport aux valeurs réelles.

plus n est grand plus n va etre fiable. Plus il y a de données, plus les bornes d'erreur vont se resserer



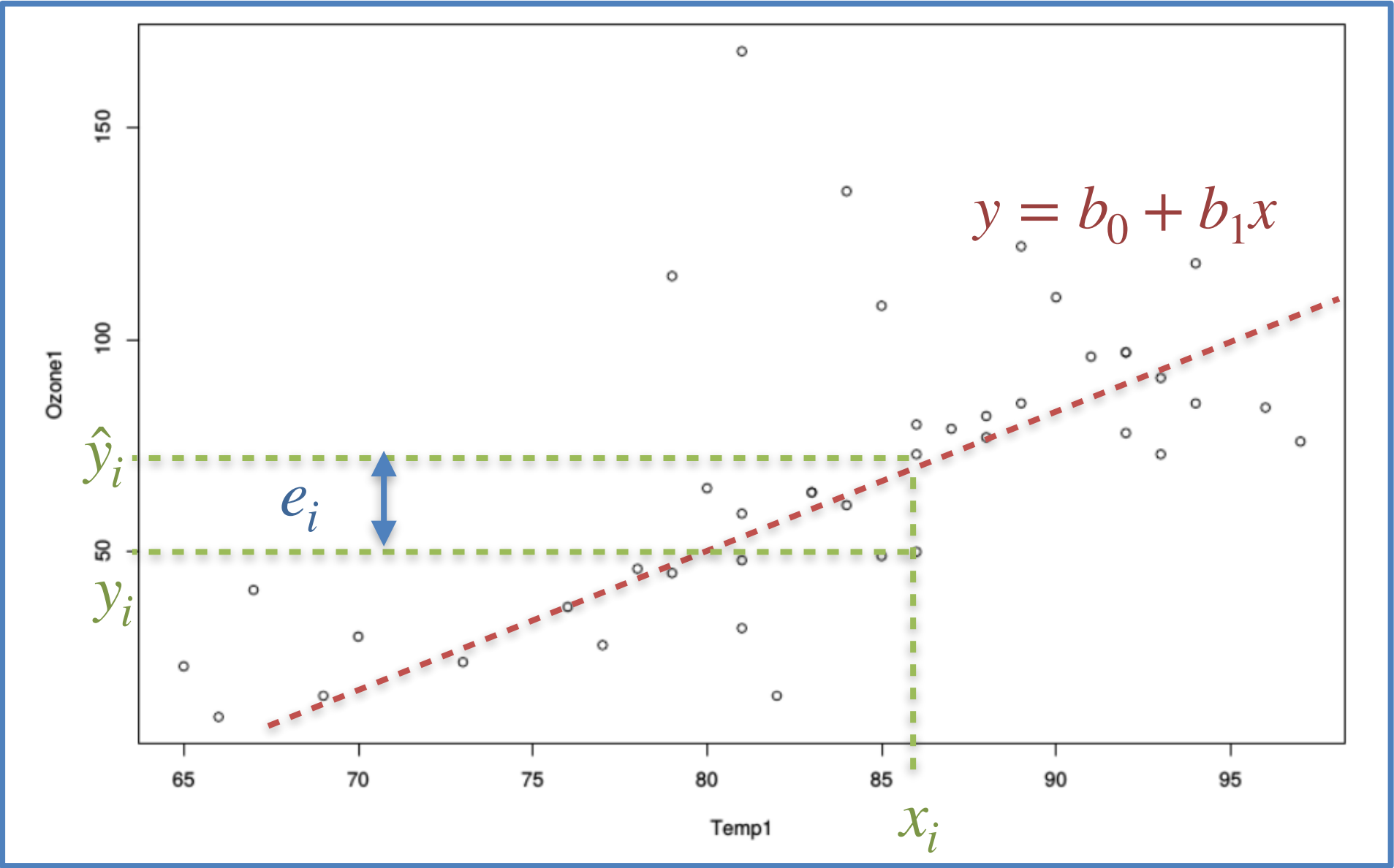
les estimateurs de  $\beta_0$  et  $\beta_1$  au sens des moindres sont :

$$b_1 = \frac{s_{xy}}{s_x^2},$$
$$b_0 = \bar{y} - b_1 \bar{x}.$$

Niveau d’incertitude lié à l’estimation d’un  $y_0$  à partir d’un  $x_0$

Enfin, connaissant une valeur  $x_0$ , on définit deux intervalles de confiance de prédiction à partir de la valeur prédite  $\hat{y}_0 = b_0 + b_1 x_0$ . Le premier encadre  $E(Y)$  sachant  $X = x_0$  ; le deuxième, encadre  $y_0$  et est plus grand car il tient compte de la variance totale  $\sigma_u^2 + Var(\hat{y}_0)$  :

$$\hat{y}_0 \pm t_{\alpha/2;(n-2)} s \left( \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{(n-1)s_x^2} \right)^{1/2},$$
$$\hat{y}_0 \pm t_{\alpha/2;(n-2)} s \left( 1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{(n-1)s_x^2} \right)^{1/2}.$$



les estimateurs de  $\beta_0$  et  $\beta_1$  au sens des moindres sont :

$$b_1 = \frac{s_{xy}}{s_x^2},$$
$$b_0 = \bar{y} - b_1 \bar{x}.$$

Qualité d’ajustement

On rappelle que la variance  $\sigma_u^2$  est estimée par la variation résiduelle :

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n e_i^2 = \frac{1}{n-1} \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \frac{1}{n-1} \sum_{i=1}^n (y_i - b_0 + b_1 x_i)^2.$$

et que :

$$s_x^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 = \frac{1}{n-1} \sum_{i=1}^n \left( x_i - \left( \frac{1}{n} \sum_{i=1}^n x_i \right) \right)^2$$

$$s_y^2 = \frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2 = \frac{1}{n-1} \sum_{i=1}^n \left( y_i - \left( \frac{1}{n} \sum_{i=1}^n y_i \right) \right)^2$$

$$s_{xy} = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{n-1} \sum_{i=1}^n \left( x_i - \left( \frac{1}{n} \sum_{i=1}^n x_i \right) \right) \left( y_i - \left( \frac{1}{n} \sum_{i=1}^n y_i \right) \right)$$

Dans l’optique de mesurer la qualité d’ajustement du modèle, il est d’usage de décomposer les sommes de carrés des écarts à la moyenne sous la forme ci-dessous :

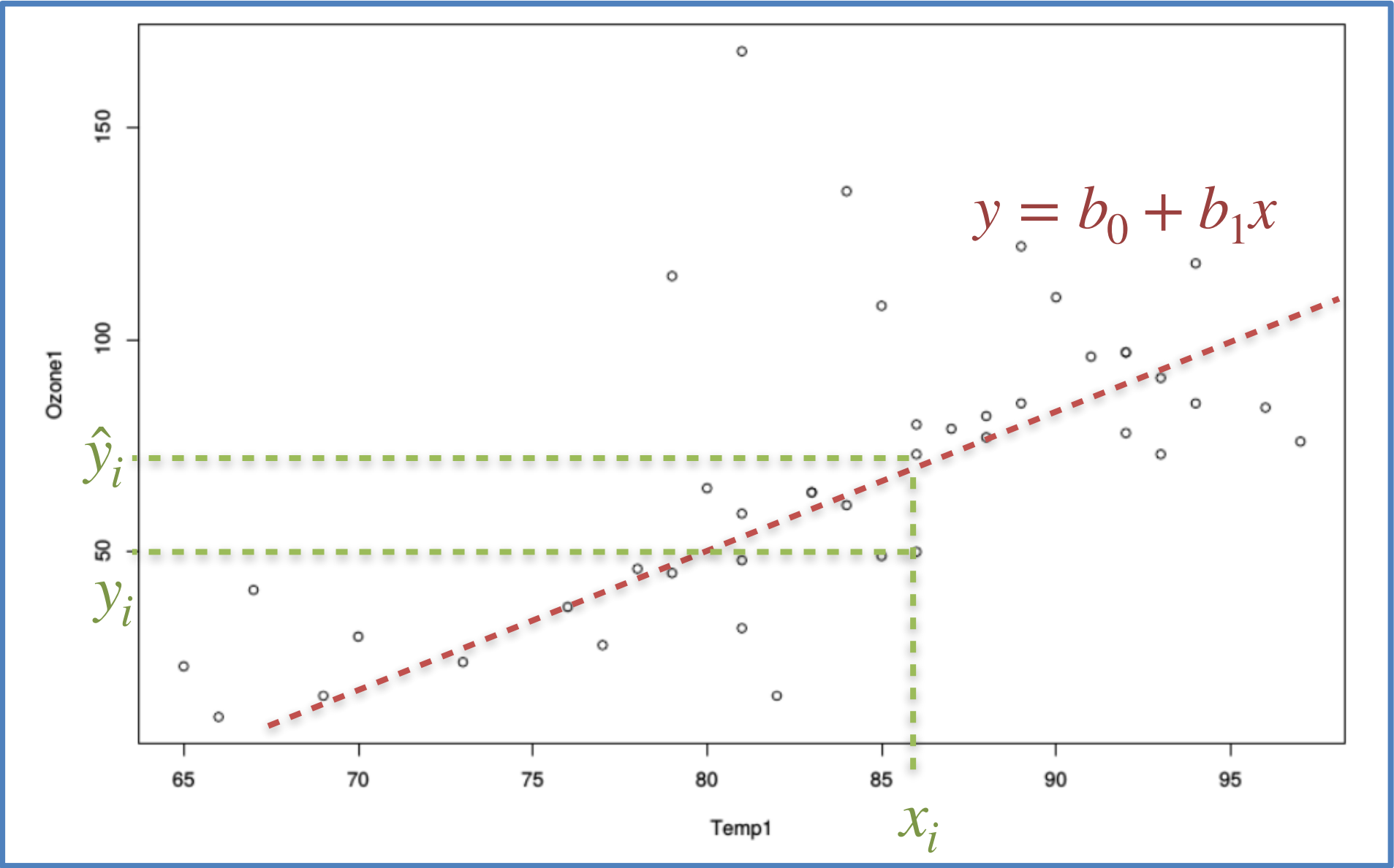
- Sum of Squares Total :  $SST = (n-1)s_y^2$
- Sum of Squares Regression :  $SSR = (n-1) \frac{s_{xy}^2}{s_x^2}$
- Sum of Squares Errors :  $SST = (n-1)s^2$

On appelle alors *coefficient de détermination* la quantité :

$$R^2 = r^2 = \frac{s_{xy}^2}{s_x^2 s_y^2} = 1 - \frac{s^2}{s_y^2} = \frac{SSR}{SST}$$

R2 : quantité de variabilité (est ce que la variabilité de mes prédictions collent avec la variabilité des données d'apprentissage)

qui exprime le rapport entre la variance expliquée par le modèle et la variance totale. En pratique, si  $R^2$  vaut par exemple 0.79, cela signifie que 79% de la variablilité de  $Y$  a été capturée par le modèle linéaire et que seulement 21% restent à expliquer.





Détection d'outliers

Effet levier

Une première indication est donnée par l'éloignement de  $x_i$  par rapport à la moyenne  $\bar{x}$ . En effet, écrivons les prédicteurs  $y_i$  comme combinaisons linéaires des observations :

$$\hat{y}_i = b_0 + b_1x_i = \sum_{j=1}^n h_{ij}y_j$$

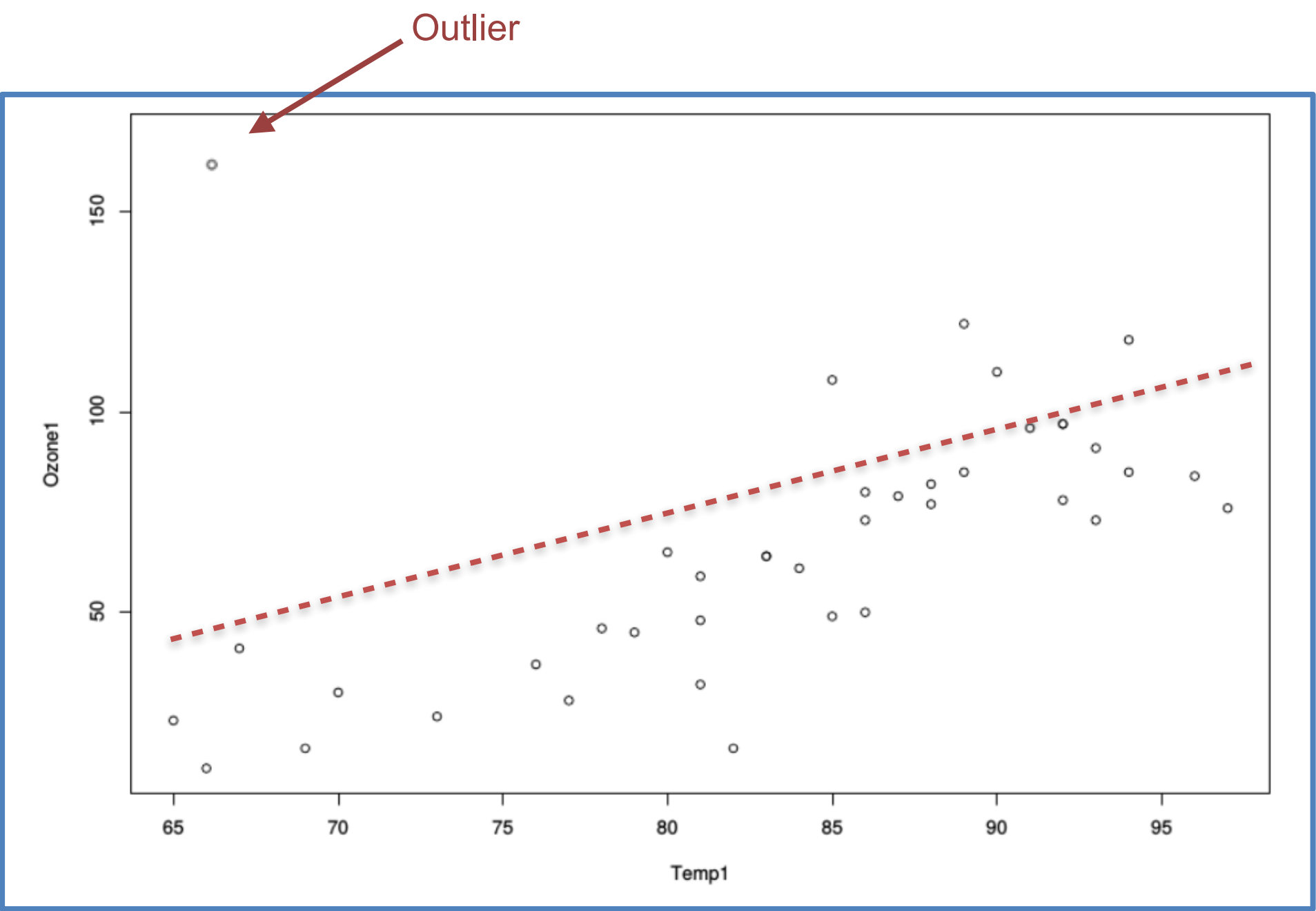
avec

$$h_{ij} = \frac{1}{n} + \frac{(x_i - \bar{x})(x_j - \bar{x})}{\sum_{k=1}^n (x_k - \bar{x})^2}$$

en notant **H** la matrice (hat matrix) des  $h_{ij}$  ceci s'exprime encore matriciellement :

$$\hat{\mathbf{y}} = \mathbf{H}\mathbf{y}$$

Les éléments diagonaux  $h_{ii}$  de cette matrice mesurent ainsi l'impact ou l'importance du rôle que joue  $y_i$  dans l'estimation de  $\hat{y}_i$ .



Détection d’outliers

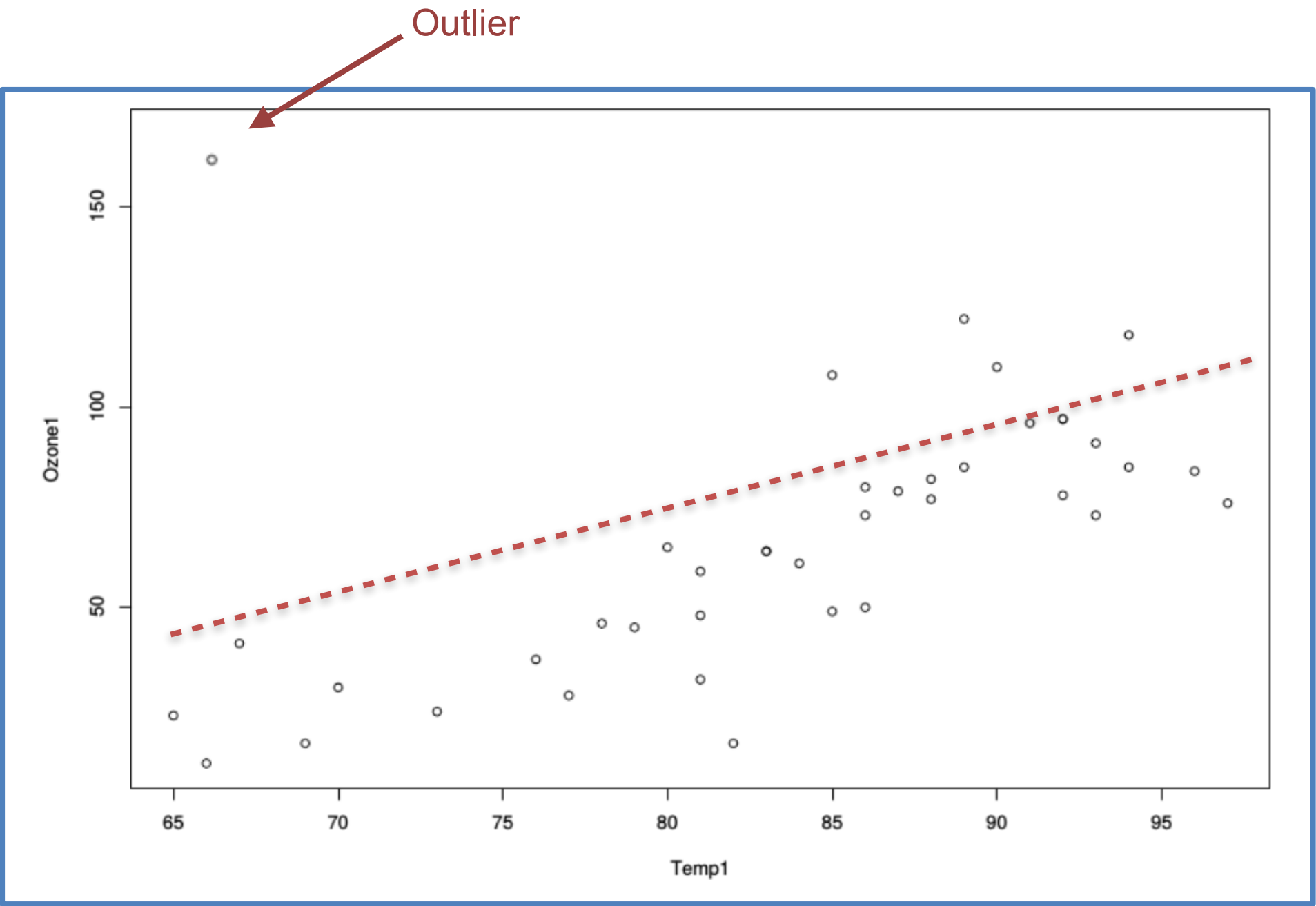
Diagnostics

Un dernier indicateur couramment utilisé est la distance de Cook :

$$D_i = \frac{\sum_{j=1}^n (\widehat{y_{(i)j}} - \widehat{y_j})^2}{2s^2} = \frac{h_{ii}}{2(1 - h_{ii})} r_i^2, \forall i$$

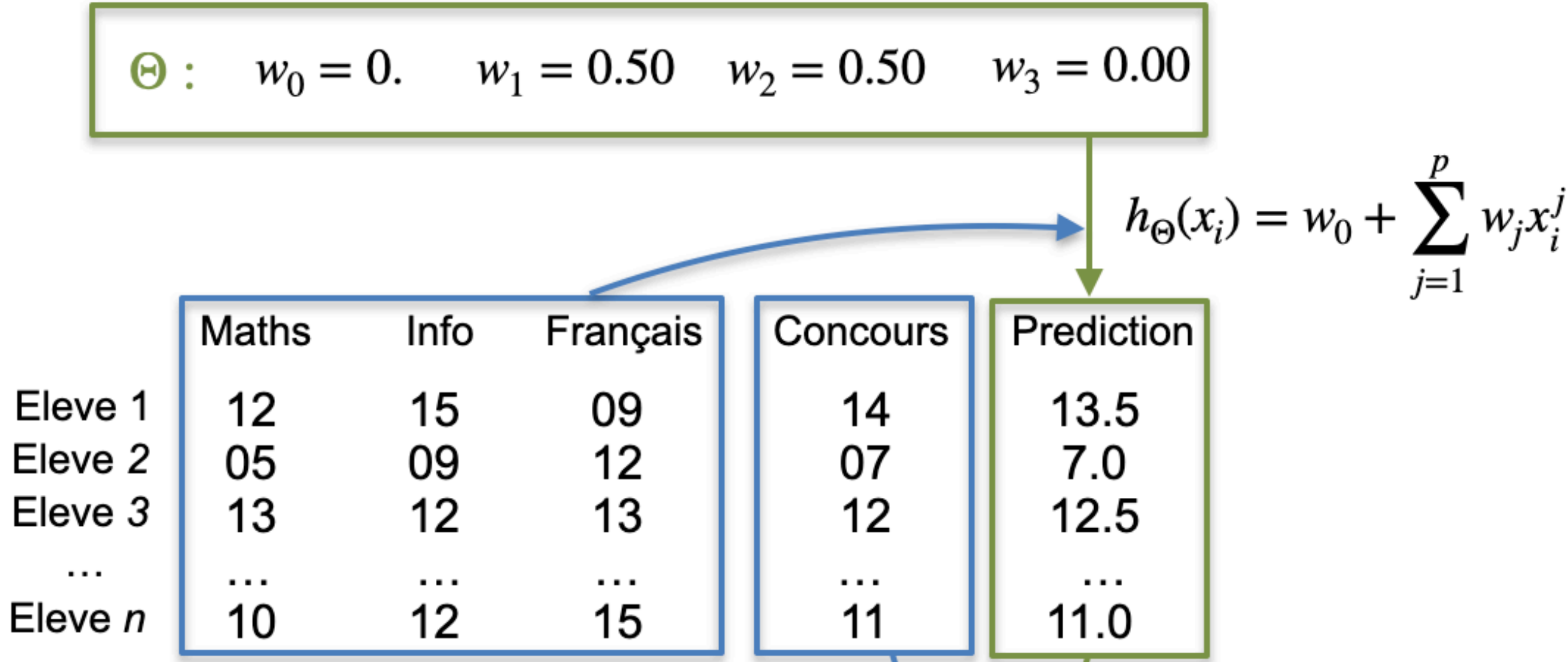
qui mesure l’influence de chaque observation  $i$  sur l’ensemble des prévisions en prenant en compte effet levier et importance des résidus.

[pour Di on enlève la prédiction du point yi](#)





Les modèles classiques de régression (linéaire, logistique) sont anciens et moins l'occasion de battage médiatique que ceux récents issus de l'apprentissage automatique. Néanmoins, ils présentent un grand intérêt compte tenu de leur robustesse, de leur stabilité face à des fluctuations d'échantillons et de leur capacité à passer à l'échelle pour des données massives. Ils restent ainsi toujours très utilisés en production notamment lorsque la fonction à modéliser est bien linéaire et qu'il serait contre productif de chercher plus compliqué.



Une variable quantitative **Y** dite à expliquer (ou encore, réponse, exogène, dépendante) est mise en relation avec  $p$  variables quantitatives  $\mathbf{X}^1, \dots, \mathbf{X}^p$  dites explicatives (ou encore de contrôle, endogènes, indépendantes, régresseurs, prédicteurs).

Les données sont supposées provenir d'un échantillon statistique de  $n$  observations, chacune étant dans  $\mathbb{R}^{(p+1)}$  (avec  $n > p + 1$ ) :

$$(x_i^1, \dots, x_i^j, \dots, x_i^p, y_i), i = 1, \dots, n$$

L'écriture du modèle linéaire dans cette situation conduit à supposer que l'espérance de **Y** appartient au sous-espace de  $\mathbb{R}^n$  engendré par  $\{\mathbf{1}, \mathbf{X}^1, \dots, \mathbf{X}^p\}$  où **1** désigne le vecteur de  $\mathbb{R}^n$  constitué de 1s. C'est-à-dire que les  $(p + 1)$  variables aléatoires vérifient :

$$Y_i = \beta_0 + \beta_1 X_i^1 + \beta_2 X_i^2 + \dots + \beta_p X_i^p + \varepsilon_i, i = 1, 2, \dots, n$$

- avec les hypothèses suivantes :
- Les  $\varepsilon_i$  sont des termes d'erreur indépendants et identiquement distribués, *i.e.*  $E(\varepsilon_i) = 0, Var(\varepsilon) = \sigma^2 \mathbf{I}$ .
  - Les termes de  $\mathbf{X}^j$ , *i.e.* du vecteur qui contient les observations de la  $j^{eme}$  variable, sont supposés déterministes (facteurs contrôlés). Dans certain contextes, on suppose alternativement que l'erreur  $\varepsilon$  est indépendante de la distribution conjointe de  $\mathbf{X}^1, \dots, \mathbf{X}^p$ . On écrit dans ce cas que  $E(\mathbf{Y} | \mathbf{X}^1, \dots, \mathbf{X}^p) = \beta_0 + \beta_1 \mathbf{X}^1 + \beta_2 \mathbf{X}^2 + \dots + \beta_p \mathbf{X}^p$  et que  $Var(Y | \mathbf{X}^1, \dots, \mathbf{X}^p) = \sigma^2$ .
  - Les paramètres inconnus  $\beta_0, \dots, \beta_p$  sont supposés constants.
  - En option, pour l'étude spécifique des lois des estimateurs, une quatrième hypothèse considère la normalité de la variable d'erreur  $\varepsilon$  (*i.e.*  $\varepsilon \sim \mathcal{N}(0, \sigma^2 \mathbf{I})$ ). Les  $\varepsilon_i$  sont alors i.i.d. de loi  $\mathcal{N}(0, \sigma^2)$ .



Les données sont rangées dans une matrice  $\mathbf{X}$  de taille  $(n \times (p + 1))$  de terme général  $X_i^j$ , dont la première colonne contient le vecteur  $\mathbf{1}$  (c'est à dire  $X_0^i = 1$ ), et dans un vecteur  $\mathbf{Y}$  de terme général  $Y_i$ . En notant les vecteurs  $\varepsilon = [\varepsilon_1 \dots \varepsilon_n]'$  et  $\beta = [\beta_0 \beta_1 \dots \beta_p]'$ , le modèle s'écrit matriciellement :

$$\mathbf{Y} = \mathbf{X}\beta + \varepsilon.$$

Ce modèle est détaillé ci-dessous :

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_1^1 & x_1^2 & \dots & x_1^p \\ 1 & x_2^1 & x_2^2 & \dots & x_2^p \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_n^1 & x_n^2 & \dots & x_n^p \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

2.2.2 Estimation

Conditionnellement à la connaissance des valeurs des  $\mathbf{X}^j$  , les paramètres inconnus du modèle, le vecteur  $\beta$  et le paramètre de nuisance  $\sigma^2$ , sont estimés par minimisation des carrés des écarts (M.C.) ou encore par maximisation de la vraisemblance (M.V.) en considérant l’hypothèse de la normalité de la variable d’erreur. Les estimateurs ont alors les mêmes expressions, l’hypothèse de normalité et l’utilisation de la vraisemblance conférant à ces derniers des propriétés complémentaires.

Etudions l’estimation par moindres carrés. L’expression à minimiser sur  $\beta \in \mathbb{R}^{p+1}$  s’écrit :

$$\sum_{i=1}^n (Y_i - \beta_0 - \beta_1 X_i^1 - \dots - \beta_p X_i^p)^2 = \| \mathbf{Y} - \mathbf{X}\beta \|^2 \tag{2.1}$$

$$= \mathbf{Y}'\mathbf{Y} - 2\beta'\mathbf{X}'\mathbf{Y} + \beta'\mathbf{X}'\mathbf{X}\beta \tag{2.2}$$

Par dérivation matricielle de la dernière équation on obtient les équations normales :

$$\mathbf{X}'\mathbf{Y} - \mathbf{X}'\mathbf{X}\beta = 0$$

2.2.2 Estimation

Conditionnellement à la connaissance des valeurs des  $\mathbf{X}^j$  , les paramètres inconnus du modèle, le vecteur  $\beta$  et le paramètre de nuisance  $\sigma^2$ , sont estimés par minimisation des carrés des écarts (M.C.) ou encore par maximisation de la vraisemblance (M.V.) en considérant l’hypothèse de la normalité de la variable d’erreur. Les estimateurs ont alors les mêmes expressions, l’hypothèse de normalité et l’utilisation de la vraisemblance conférant à ces derniers des propriétés complémentaires.

Etudions l’estimation par moindres carrés. L’expression à minimiser sur  $\beta \in \mathbb{R}^{p+1}$  s’écrit :

$$\sum_{i=1}^n (Y_i - \beta_0 - \beta_1 X_i^1 - \dots - \beta_p X_i^p)^2 = \| \mathbf{Y} - \mathbf{X}\beta \|^2 \tag{2.1}$$

$$= \mathbf{Y}'\mathbf{Y} - 2\beta'\mathbf{X}'\mathbf{Y} + \beta'\mathbf{X}'\mathbf{X}\beta \tag{2.2}$$

Par dérivation matricielle de la dernière équation on obtient les équations normales :

$$\mathbf{X}'\mathbf{Y} - \mathbf{X}'\mathbf{X}\beta = 0$$

Nous faisons l’hypothèse supplémentaire que la matrice  $\mathbf{X}'\mathbf{X}$  est inversible, c’est-à-dire que la matrice  $\mathbf{X}$  est de rang  $(p + 1)$  et donc qu’il n’existe pas de colinéarité entre ses colonnes. Si cette hypothèse n’est pas vérifiée, il suffit en principe de supprimer des colonnes de  $\mathbf{X}$  et donc des variables du modèle. Une approche de réduction de dimension (régression ridge, Lasso, PLS...) est en pratique à mettre en oeuvre (voir Section 3.3). Alors, l’estimation des paramètres  $\beta_j$  est donnée par :

$$\widehat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$$



### 2.2.3 Prédiction

Connaissant les valeurs des variables  $\mathbf{X}^j$  pour une nouvelle observation :  $x_0 = [x_0^1, x_0^2, \dots, x_0^p]'$  appartenant au domaine dans lequel l'hypothèse de linéarité reste valide, une prédiction, notée  $\hat{y}_0$  de  $\mathbf{Y}$  ou  $E(\mathbf{Y})$  est donnée par :

$$\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0^1 + \dots + \hat{\beta}_p x_0^p.$$

### 2.2.3 Prévion

Connaissant les valeurs des variables  $\mathbf{X}^j$  pour une nouvelle observation :  $x_0 = [x_0^1, x_0^2, \dots, x_0^p]'$  appartenant au domaine dans lequel l'hypothèse de linéarité reste valide, une prévision, notée  $\hat{y}_0$  de  $\mathbf{Y}$  ou  $E(\mathbf{Y})$  est donnée par :

$$\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0^1 + \dots + \hat{\beta}_p x_0^p.$$

Les intervalles de confiance des prévisions de  $\mathbf{Y}$  et  $E(\mathbf{Y})$ , pour une valeur  $\mathbf{x}_0 \in \mathbb{R}^p$  et en posant  $\mathbf{v}_0 = (1|\mathbf{x}_0')' \in \mathbb{R}^{p+1}$ , sont respectivement

$$\begin{aligned} \hat{y}_0 &\pm t_{\alpha/2; (n-p-1)} \hat{\sigma} (1 + \mathbf{v}_0' (\mathbf{X}'\mathbf{X})^{-1} \mathbf{v}_0)^{1/2}, \\ \hat{y}_0 &\pm t_{\alpha/2; (n-p-1)} \hat{\sigma} (\mathbf{v}_0' (\mathbf{X}'\mathbf{X})^{-1} \mathbf{v}_0)^{1/2}. \end{aligned}$$

Il est intéressant de remarquer que ces intervalles de confiance dépendent d'une loi de Student à  $n-p-1$  degrés de liberté (voir appendice A). A dimension des observations fixée  $p$ , plus  $n$  est grand, plus les valeurs de la loi de Student seront faible, et ainsi les marges seront réduites. Cependant, plus  $p$  est proche de  $n$ , plus les marges sont élevées. En parallèle, les variances de ces prévisions, comme celles des estimations des paramètres, dépendent aussi directement du conditionnement de la matrice  $\mathbf{X}'\mathbf{X}$  de par le terme  $\mathbf{v}_0' (\mathbf{X}'\mathbf{X})^{-1} \mathbf{v}_0$ .