12. Régression linéaire simple

MTH2302D

S. Le Digabel, École Polytechnique de Montréal

A2017

(v2)

Plan

Régression linéaire : introduction

But : établir un lien entre une variable dépendante Y et une variable indépendante X pour pouvoir ensuite faire des prévisions sur Y lorsque X est mesurée.

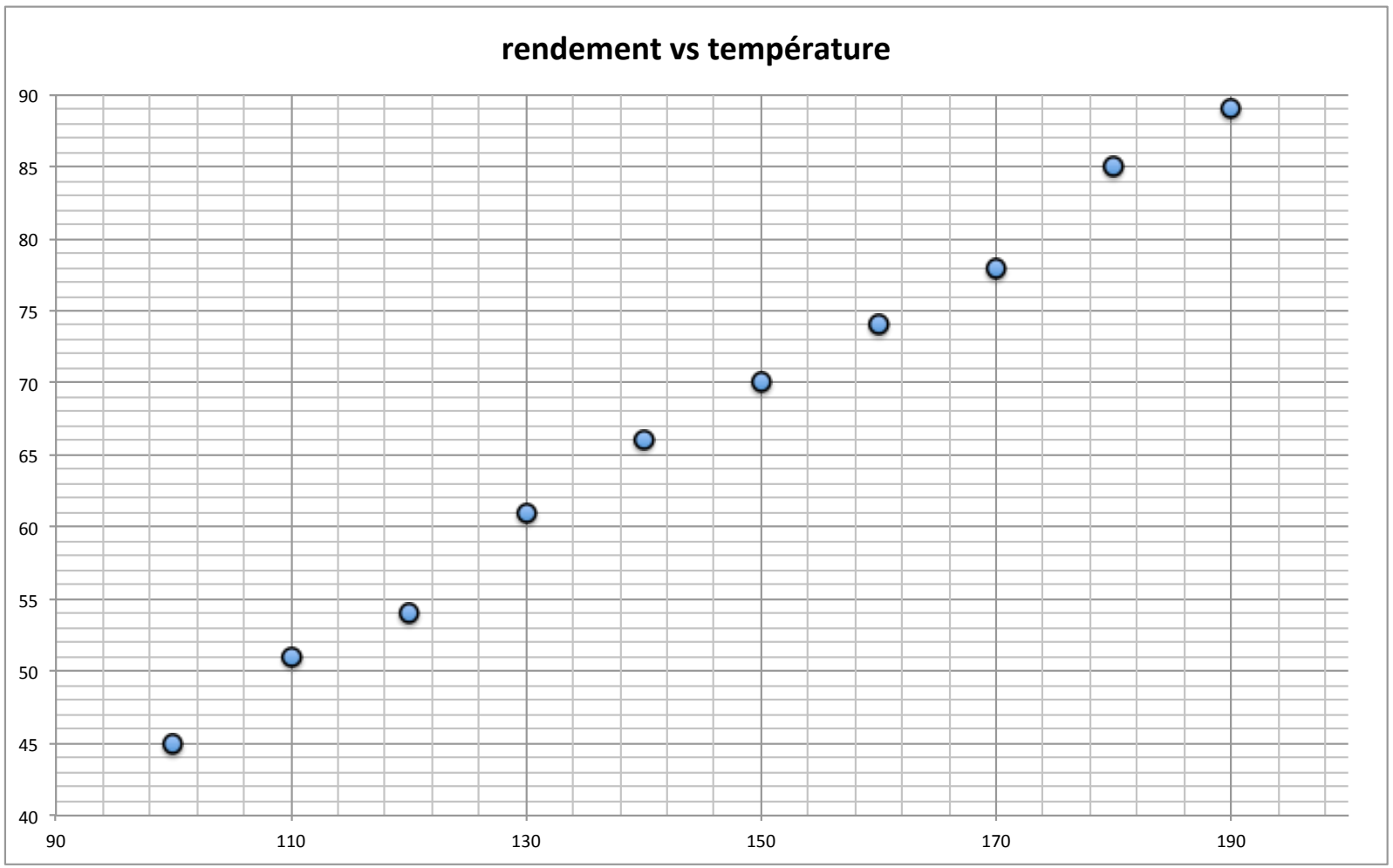
Exemple 1

L'analyse de la température de fonctionnement d'un procédé chimique sur le rendement du produit a donné les valeurs suivantes pour la température X_i et le rendement correspondant Y_i :

Température °C	Rendement %	Température °C	Rendement %
100	45	150	70
110	51	160	74
120	54	170	78
130	61	180	85
140	66	190	89

Exemple 1 (suite)

Le graphe ci-dessous représente les points (X_i, Y_i) pour ces données et suggère une relation linéaire entre X et Y.



Modèle linéaire

Définition

Un *modèle de régression linéaire simple* est de la forme

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

où

- Y est la *variable dépendante* (une v.a.).
- β_0 et β_1 sont les *coefficients* (ordonnée à l'origine et pente).
- X est la *variable indépendante* (variable explicative).
- ε est une *erreur* aléatoire.

Modèle linéaire (suite)

L'espérance de Y pour chaque X est le point sur la droite d'équation $\mathsf{E}(Y|X) = \beta_0 + \beta_1 X$.

On suppose que

- Pour chaque valeur de X, $\mathsf{E}(\varepsilon) = 0$ et $\mathsf{V}(\varepsilon) = \sigma^2$.
- $\varepsilon \sim \mathsf{N}(0, \sigma^2)$.
- Les erreurs ε sont indépendantes (non corrélées).

On cherche à

- Estimer les paramètres β_0, β_1 et σ^2.
- Vérifier si le modèle est adéquat.

Paramètres β_0 et β_1

Supposons que n paires d'observations (X_1, Y_1), (X_2, Y_2), ..., (X_n, Y_n) ont été faites. Substituant dans le modèle linéaire, on obtient

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \quad \Rightarrow \quad \varepsilon_i = Y_i - \beta_0 - \beta_1 X_i.$$

Les coefficients sont déterminés par la méthode des moindres carrés qui minimise la somme des carrés des erreurs :

$$L(\beta_0, \beta_1) = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2.$$

On résout le système de deux équations à deux inconnues $\nabla L(\hat{\beta}_0, \hat{\beta}_1) = 0$.

Paramètres β_0 et β_1 (suite)

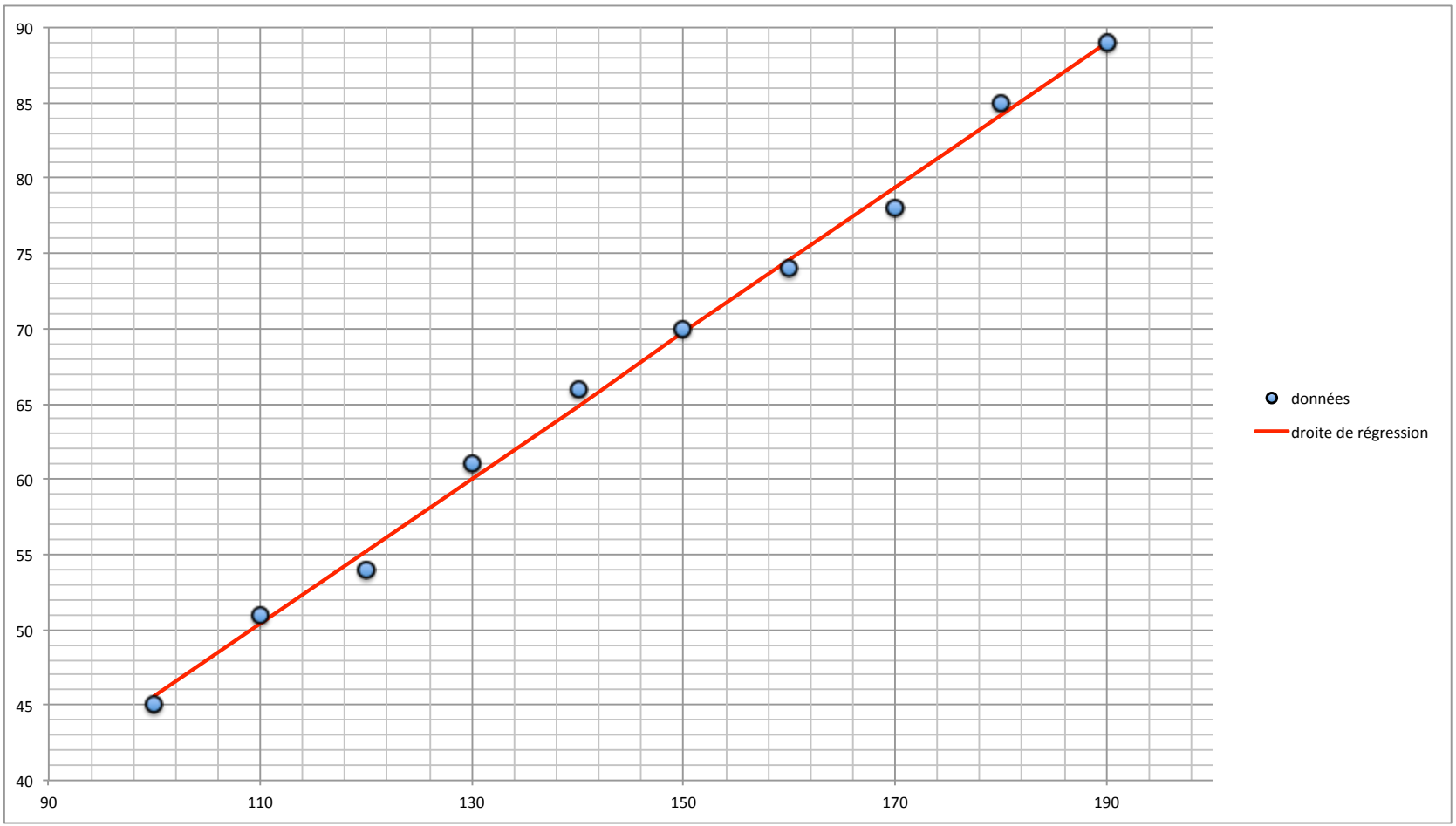
$$\nabla L(\hat{\beta}_0, \hat{\beta}_1) = 0 \Rightarrow \begin{cases} \hat{\beta}_0 &= \overline{Y} - \hat{\beta}_1 \overline{X} \\ \hat{\beta}_1 &= \dfrac{\sum_{i=1}^n X_i Y_i - n\ \overline{X}\ \overline{Y}}{\sum_{i=1}^n X_i^2 - n\overline{X}^2} = \dfrac{S_{XY}}{S_{XX}} \end{cases}$$

avec

- $\overline{X} = \frac{1}{n}\sum_{i=1}^n X_i$ et $\overline{Y} = \frac{1}{n}\sum_{i=1}^n Y_i$.
- $S_{XX} = \sum_{i=1}^n (X_i - \overline{X})^2 = \sum_{i=1}^n X_i^2 - n\overline{X}^2 = (n-1)S^2$.
- $S_{YY} = \sum_{i=1}^n (Y_i - \overline{Y})^2 = \sum_{i=1}^n Y_i^2 - n\overline{Y}^2$.
- $S_{XY} = \sum_{i=1}^n (X_i - \overline{X})(Y_i - \overline{Y}) = \sum_{i=1}^n X_i Y_i - n\ \overline{X}\ \overline{Y}$.

Exemple 2 : retrouver ces formules.

Droite de régression pour l'exemple 1



Voir fichier Excel.

Point de vue algébrique

- Étant donnés n *points de données* $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$ de $\mathbb{R}^2$, on essaie de trouver l'équation d'une droite qui passe par les n points.
- Cette équation est $Y = \beta_0 + \beta_1 X$ avec $\beta_0, \beta_1 \in \mathbb{R}$.
- β_0 et β_1 devraient être les solutions du système $A\mathbf{x} = \mathbf{b}$ avec

$$A = \begin{bmatrix} 1 & X_1 \\ 1 & X_2 \\ \vdots & \vdots \\ 1 & X_n \end{bmatrix}, \ \mathbf{x} = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}, \ \mathbf{b} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} .$$

- Résolution au sens des moindres carrés :

$$(\hat{\beta}_0, \hat{\beta}_1) = \left(A^\top A\right)^{-1} A^\top \mathbf{b} \ .$$

Propriétés de β_0 et β_1

La droite de régression estimée est $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$.

Les variables aléatoires $\hat{\beta}_0$ et $\hat{\beta}_1$ sont des estimateurs de l'ordonnée à l'origine β_0 et de la pente β_1.

Théorème

1. $\mathsf{E}(\hat{\beta}_0) = \beta_0$ et $\mathsf{E}(\hat{\beta}_1) = \beta_1$ (estimateurs non biaisés).
2. $\mathsf{V}(\hat{\beta}_0) = \sigma^2 \left[\frac{1}{n} + \frac{\overline{X}^2}{S_{XX}} \right]$ et $\mathsf{V}(\hat{\beta}_1) = \frac{\sigma^2}{S_{XX}}$.
3. $\mathsf{Cov}(\hat{\beta}_0, \hat{\beta}_1) = -\frac{\sigma^2 \overline{X}}{S_{XX}}$.

Paramètre σ^2

Rappel : le modèle de régression est $Y = \beta_0 + \beta_1 X + \varepsilon$ avec $\varepsilon \sim \mathrm{N}(0, \sigma^2)$.

La différence entre la valeur estimée $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$ et la valeur observée Y_i est appelée *résidu* et est dénotée $E_i = \hat{Y}_i - Y_i$.

On définit

- La *somme des carrés dûe à l'erreur* par

$$SS_E = \sum_{i=1}^{n} E_i^2 = \sum_{i=1}^{n} (\hat{Y}_i - Y_i)^2.$$

- La *somme des carrés dûe à la régression* par

$$SS_R = \sum_{i=1}^{n} (\hat{Y}_i - \overline{Y})^2 = \hat{\beta}_1^2 S_{XX} = \frac{S_{XY}^2}{S_{XX}}.$$

Paramètre σ^2 (suite)

La quantité S_{YY} représente la variabilité totale des Y_i. On peut la décomposer par

$$S_{YY} = SS_T = SS_E + SS_R \ .$$

Théorème

1. $\mathrm{E}(SS_E) = (n-2)\sigma^2$.

2. $\hat{\sigma}^2 = \dfrac{SS_E}{n-2} \equiv MS_E$ est donc un estimateur sans biais de σ^2.

Exemple 1 (suite)

L'analyse de la température de fonctionnement d'un procédé chimique sur le rendement du produit a donné les valeurs suivantes pour la température X_i et le rendement correspondant Y_i :

Température °C	Rendement %	Température °C	Rendement %
100	45	150	70
110	51	160	74
120	54	170	78
130	61	180	85
140	66	190	89

Voir fichier Excel.

Distributions pour $\hat{\beta}_0$ et $\hat{\beta}_1$

Théorème

La statistique

$$\frac{\hat{\beta}_0 - \beta_0}{\sqrt{MS_E \left[\frac{1}{n} + \frac{\overline{X}^2}{S_{XX}}\right]}}$$

suit une loi de Student à $n-2$ degrés de liberté.

Théorème

La statistique

$$\frac{\hat{\beta}_1 - \beta_1}{\sqrt{MS_E/S_{XX}}}$$

suit une loi de Student à $n-2$ degrés de liberté.

Intervalles de confiance pour β_0 et β_1

Théorème

Intervalles de confiance bilatéraux au niveau de confiance $1-\alpha$ pour β_0 et β_1 :

$$\beta_0 = \hat{\beta}_0 \pm t_{\alpha/2;n-2}\sqrt{MS_E\left[\frac{1}{n}+\frac{\overline{X}^2}{S_{XX}}\right]}$$

$$\beta_1 = \hat{\beta}_1 \pm t_{\alpha/2;n-2}\sqrt{\frac{MS_E}{S_{XX}}}\ .$$

Voir fichier Excel.

Intervalles de confiance pour la droite de régression

Il s'agit d'un intervalle de confiance pour $\mathsf{E}(Y_0|x_0)$, la réponse moyenne à la valeur x_0.

Pour x_0 donné soit $\hat{Y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0$ l'estimateur de $\mathsf{E}(Y_0|x_0)$.

Théorème

Intervalle de confiance pour $\mathsf{E}(Y_0|x_0)$ au niveau de confiance $1-\alpha$:

$$\mathsf{E}(Y_0|x_0) = \hat{Y}_0 \pm t_{\alpha/2;n-2}\sqrt{MS_E\left[\frac{1}{n} + \frac{(\overline{X}-x_0)^2}{S_{XX}}\right]}$$

Exemple 1 (suite)

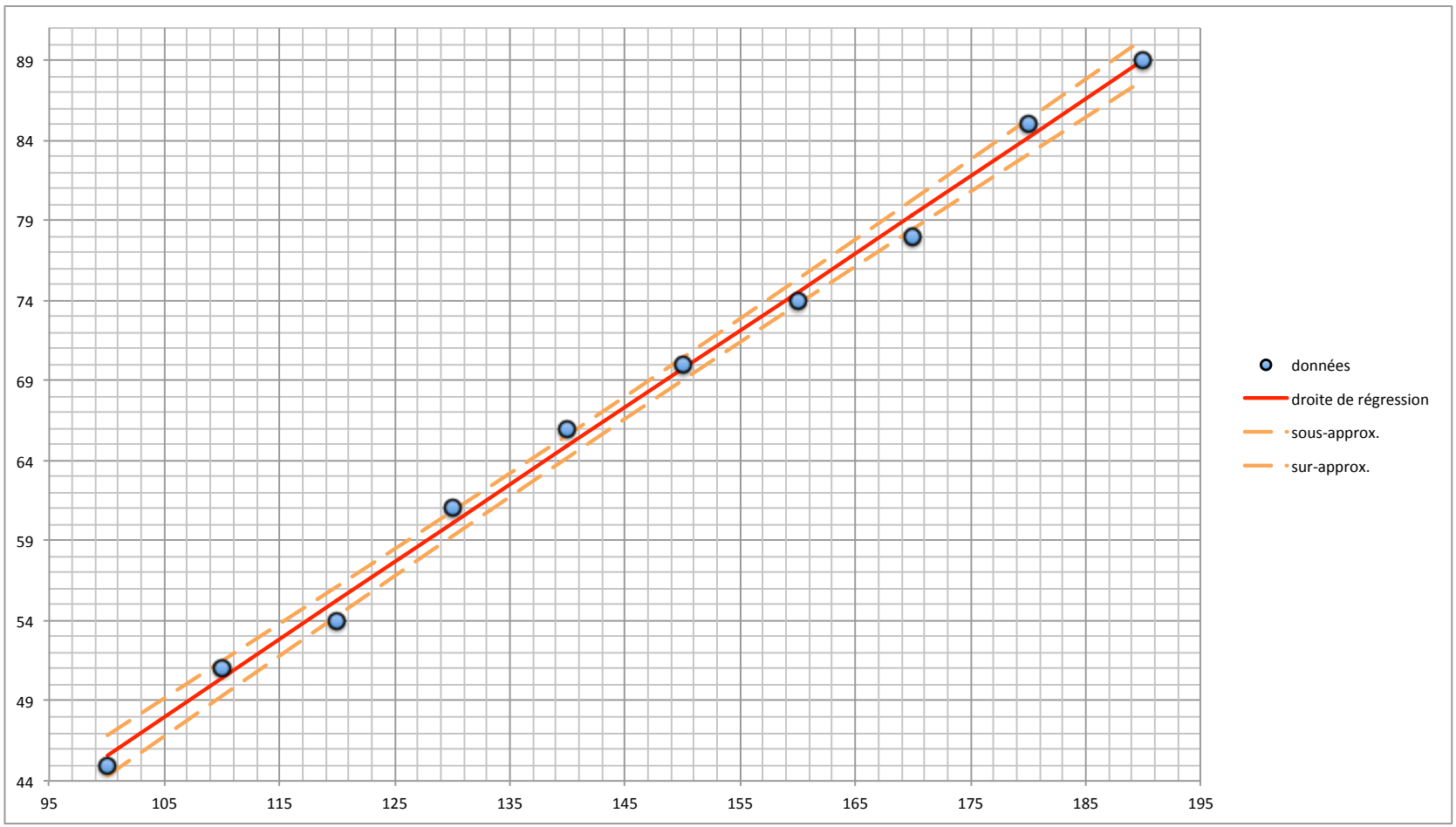
Le calcul de l'intervalle de confiance à 95% en chaque point $x_0 = X_i$, $i = 1, 2, \ldots, 10$ donne le tableau suivant :

x_0	100	110	120	130	140
$\hat{y}_0$	45.56	50.39	55.22	60.05	64.88
limites	± 1.30	± 1.10	± 0.93	± 0.79	± 0.71
x_0	150	160	170	180	190
$\hat{y}_0$	69.72	74.55	79.38	84.21	89.04
limites	± 0.71	± 0.79	0.93	± 1.10	± 1.30

Voir fichier Excel.

Exemple 1 (suite)

à partir des données du tableau précédent, on a tracé l'intervalle de confiance pour la droite de régression :



Intervalles de prévision

Soit x_0 une valeur quelconque. La valeur correspondante de Y est $Y_0 = Y|x_0 = \beta_0 + \beta_1 x_0 + \varepsilon_0$. On estime ponctuellement Y_0 par $\hat{Y}_0 = \hat{\beta}_0 \ + \hat{\beta}_1 x_0$.

La statistique

$$\frac{Y_0 - \hat{Y}_0}{\sqrt{MS_E \left[1 + \frac{1}{n} + \frac{(\overline{X} - x_0)^2}{S_{XX}}\right]}}$$

suit une loi de Student à $n-2$ degrés de liberté.

Théorème

Intervalle de prévision pour la valeur de Y en x_0 :

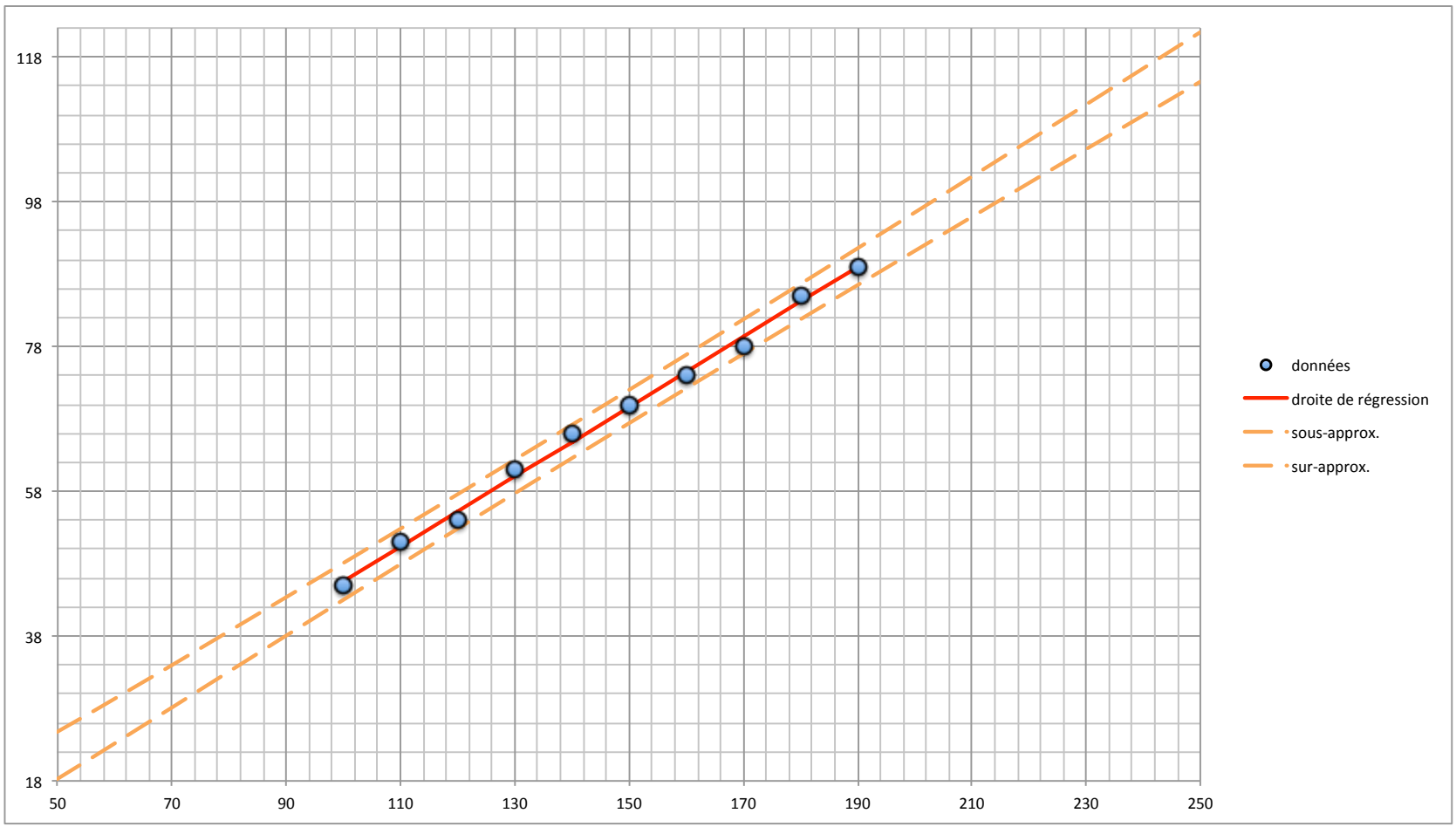
$$Y_0 = \hat{Y}_0 \pm t_{\alpha/2;n-2} \sqrt{MS_E \left[1 + \frac{1}{n} + \frac{(\overline{X} - x_0)^2}{S_{XX}}\right]}.$$

Remarques : IC vs IP

- Les longueurs des deux types d'intervalles croissent lorsque x_0 s'éloigne de $\overline{X}$.
- L'IC de la droite de régression ne convient pas pour effectuer des prévisions puisqu'il concerne la vraie réponse moyenne au point $X = x_0$, soit un paramètre de la population, et non une nouvelle observation, i.e. une nouvelle valeur pour la v.a. Y.
- L'IP en x_0 est toujours plus grand que l'IC en x_0 car il dépend de l'erreur associée aux futures observations.
- L'IP prend en compte une nouvelle observation, d'où une augmentation de $\sigma^2 \simeq MS_E$ de la variance.
- L'IP n'est valide que pour **une** nouvelle observation à la fois. Pour une série de nouvelles observations, il faut mettre à jour le modèle au fur et à mesure.
- Voir fichier Excel.

Exemple 1 (suite)

à partir des données du tableau précédent, on a tracé l'intervalle de prévision pour $\alpha = 5\%$:



Tests d'hypothèses pour β_0

La distribution

$$t_0 = \frac{\hat{\beta}_0 - \beta_{0,0}}{\sqrt{MS_E \left[\frac{1}{n} + \frac{\overline{X}^2}{S_{XX}}\right]}} \sim T_{n-2}$$

permet de tester des hypothèses du type

$$\begin{aligned} H_0 : &\quad \beta_0 = \beta_{0,0} \\ H_1 : &\quad \beta_0 \neq \beta_{0,0} \end{aligned}$$

On rejette H_0 au seuil α si $|t_0| > t_{\alpha/2;n-2}$.

Tests d'hypothèses pour β_1

La distribution

$$t_0 = \frac{\hat{\beta}_1 - \beta_{1,0}}{\sqrt{MS_E/S_{XX}}} \sim T_{n-2}$$

permet de tester des hypothèses du type

$$\begin{aligned} H_0 : &\quad \beta_1 = \beta_{1,0} \\ H_1 : &\quad \beta_1 \neq \beta_{1,0} \end{aligned}$$

On rejette H_0 au seuil α si $|t_0| > t_{\alpha/2;n-2}$.

Tableau d'analyse de la variance

L'information donnée par les valeurs S_{YY}, SS_E et SS_R est présentée dans un *tableau d'analyse de la variance* :

Source de variation	Somme des carrés	Nombre de d.d.l.	Moyenne des carrés	F_0
Régression	SS_R	1	$MS_R = \dfrac{SS_R}{1}$	$\dfrac{MS_R}{MS_E}$
Résidus	SS_E	$n-2$	$MS_E = \dfrac{SS_E}{n-2}$	
Total	$SS_T = S_{YY}$	$n-1$		

Signification de la régression

Il s'agit de tester les hypothèses

$$\begin{aligned} H_0 : &\quad \beta_1 = 0 \\ H_1 : &\quad \beta_1 \neq 0 \end{aligned}$$

Accepter H_0 implique que l'on conclut qu'il n'y a pas de relation linéaire entre X et Y. Ceci peut signifier que

- La relation entre X et Y n'est pas linéaire.
- La variation de X influe peu ou pas sur la variation de Y.

Au contraire, rejeter H_0 implique que l'on conclut que la variation de X influe sur la variation de Y.

Le critère est : rejeter H_0 au seuil α si $F_0 > F_{\alpha;1,n-2}$, ou encore si la valeur-P calculée est petite, avec valeur-P=$P(F_{1,n-2} \geq F_0)$.

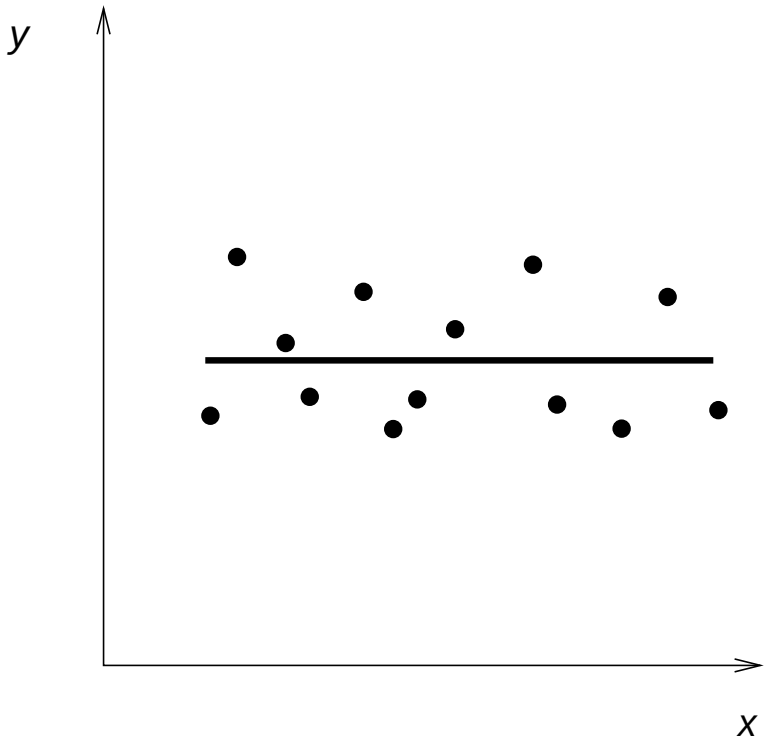
Exemple 1 : tableau d'analyse de la variance

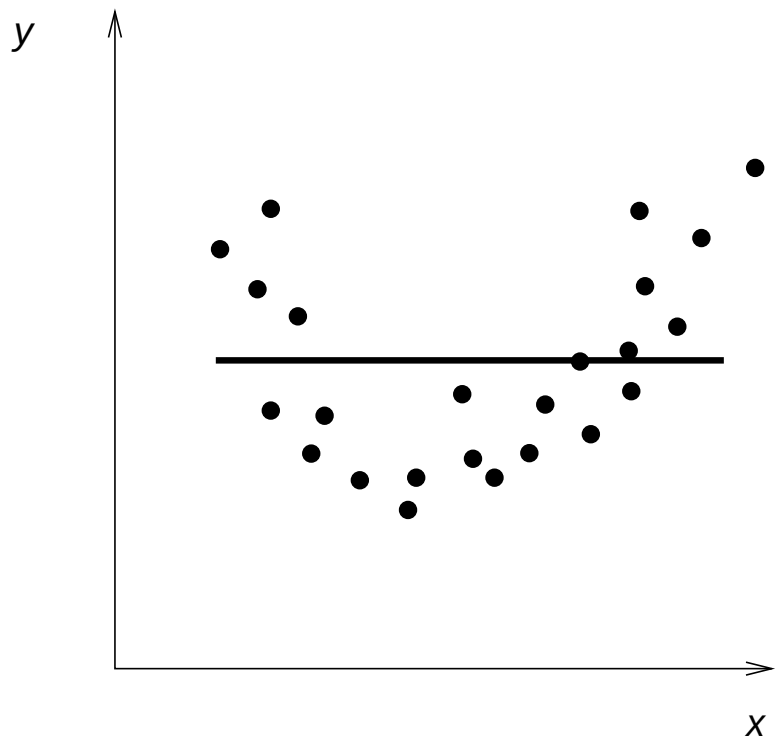
Source de variation	Somme des carrés	Nombre de d.d.l.	Moyenne des carrés	F_0
Régression	$SS_R = 1924.88$	1	$MS_R = 1924.88$	2131.57
Résidus	$SS_E = 7.22$	8	$MS_E = 0.90$	
Total	$SS_T = 1932.10$	9		

P-val. : $P(F_{1,8} \geq F_0) \simeq 5.35 \times 10^{-11} < \alpha = 5\% \Rightarrow$ on rejette H_0.

Signification de la régression (suite)

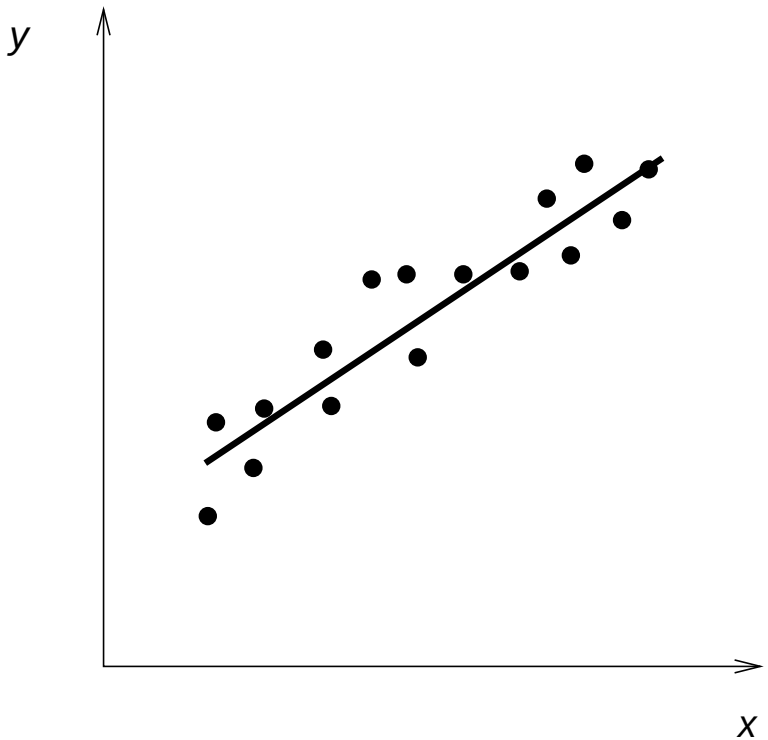
On ne rejette pas H_0 :

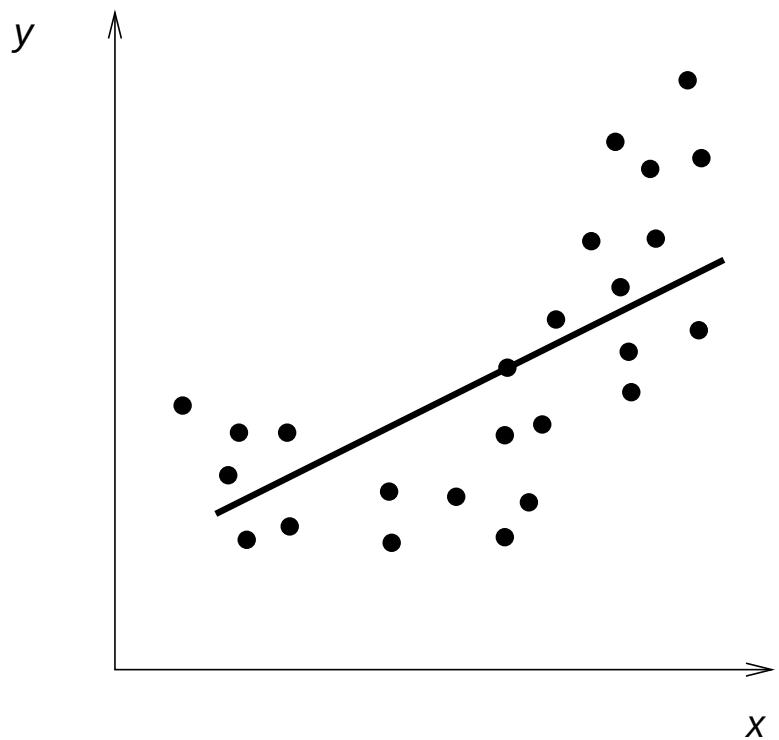




Signification de la régression (suite)

On rejette H_0 :





Rappel des hypothèses pour la régression linéaire

Tout ce qui a été fait jusqu'ici suppose que

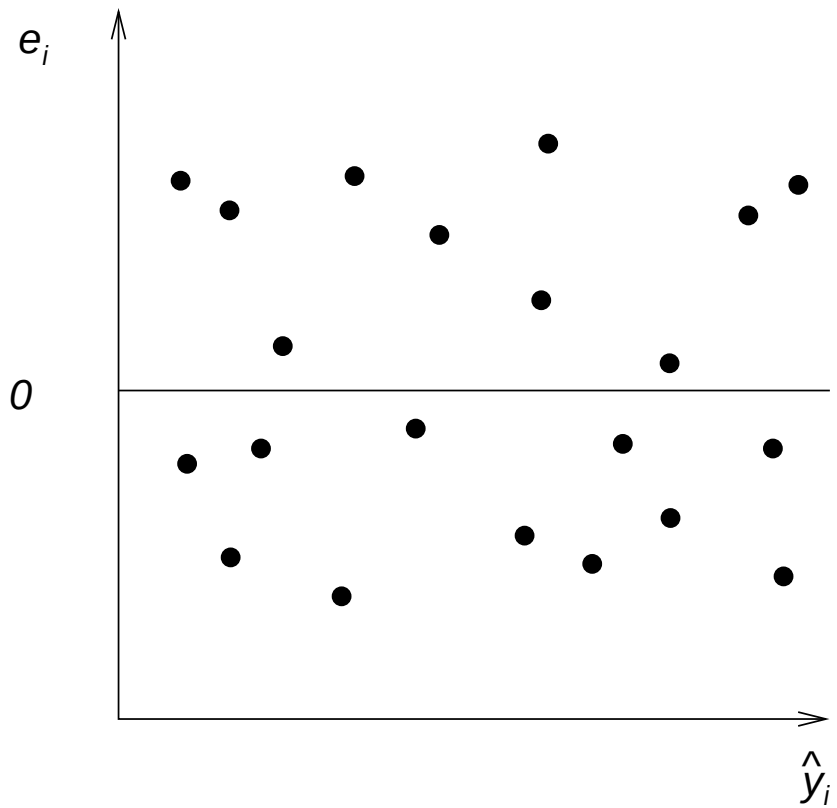
- Pour chaque X, $\mathrm{E}(\varepsilon) = 0$ et $\mathrm{V}(\varepsilon) = \sigma^2$ est constante.
- Les erreurs ε sont non corrélées.
- Les erreurs ε sont distribuées normalement.

On veut vérifier, après que les observations soient faites, si ces hypothèses sont satisfaites.

Analyse graphique des résidus

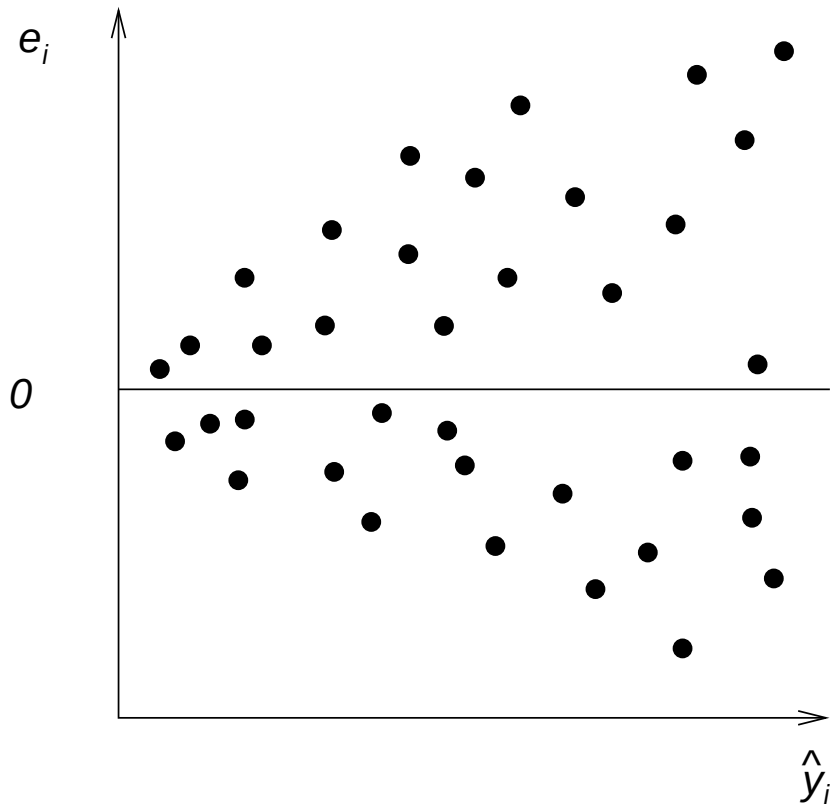
Pour vérifier l'hypothèse sur σ^2, on peut tracer le graphe des points $(\hat{Y}_i, E_i)$ ou (X_i, E_i). Les situations possibles sont illustrées ci-dessous.

Situation a) Convenable :



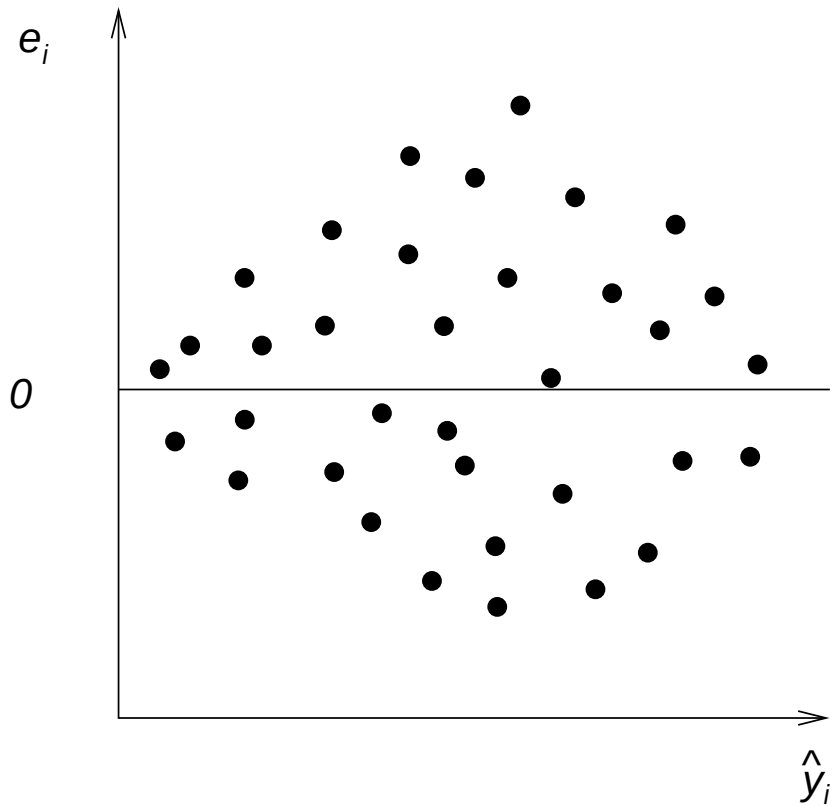
Analyse graphique des résidus (suite)

Situation b) La variance augmente avec la valeur de $\hat{Y}_i$ (ou X_i), donc σ^2 n'est pas constante :



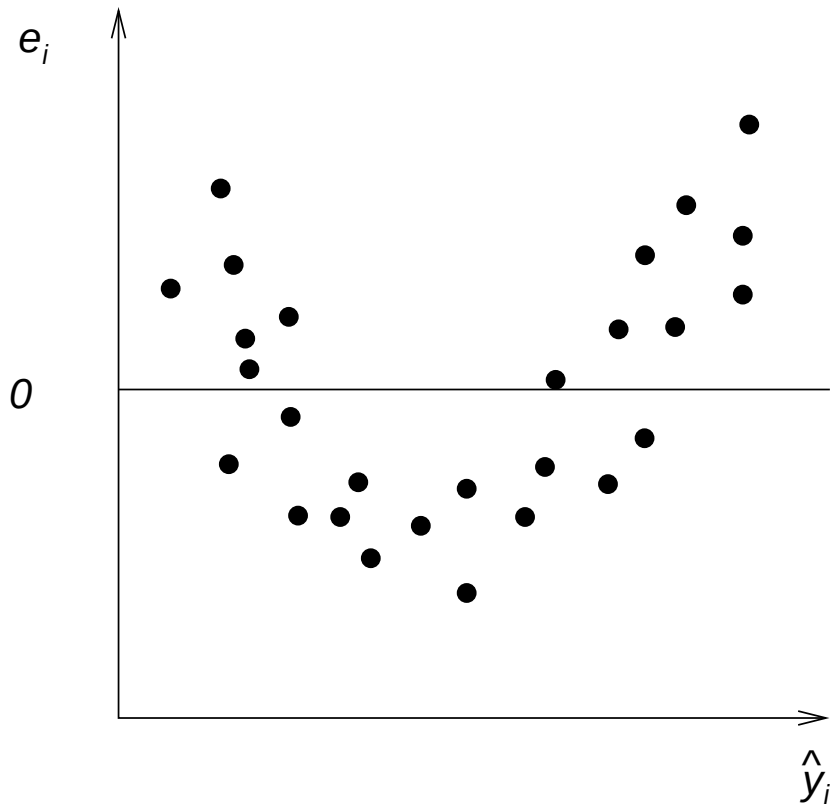
Analyse graphique des résidus (suite)

Situation c) La variance σ^2 n'est pas constante :



Analyse graphique des résidus (suite)

Situation d) Le modèle linéaire n'est pas approprié :



Test de la normalité des résidus

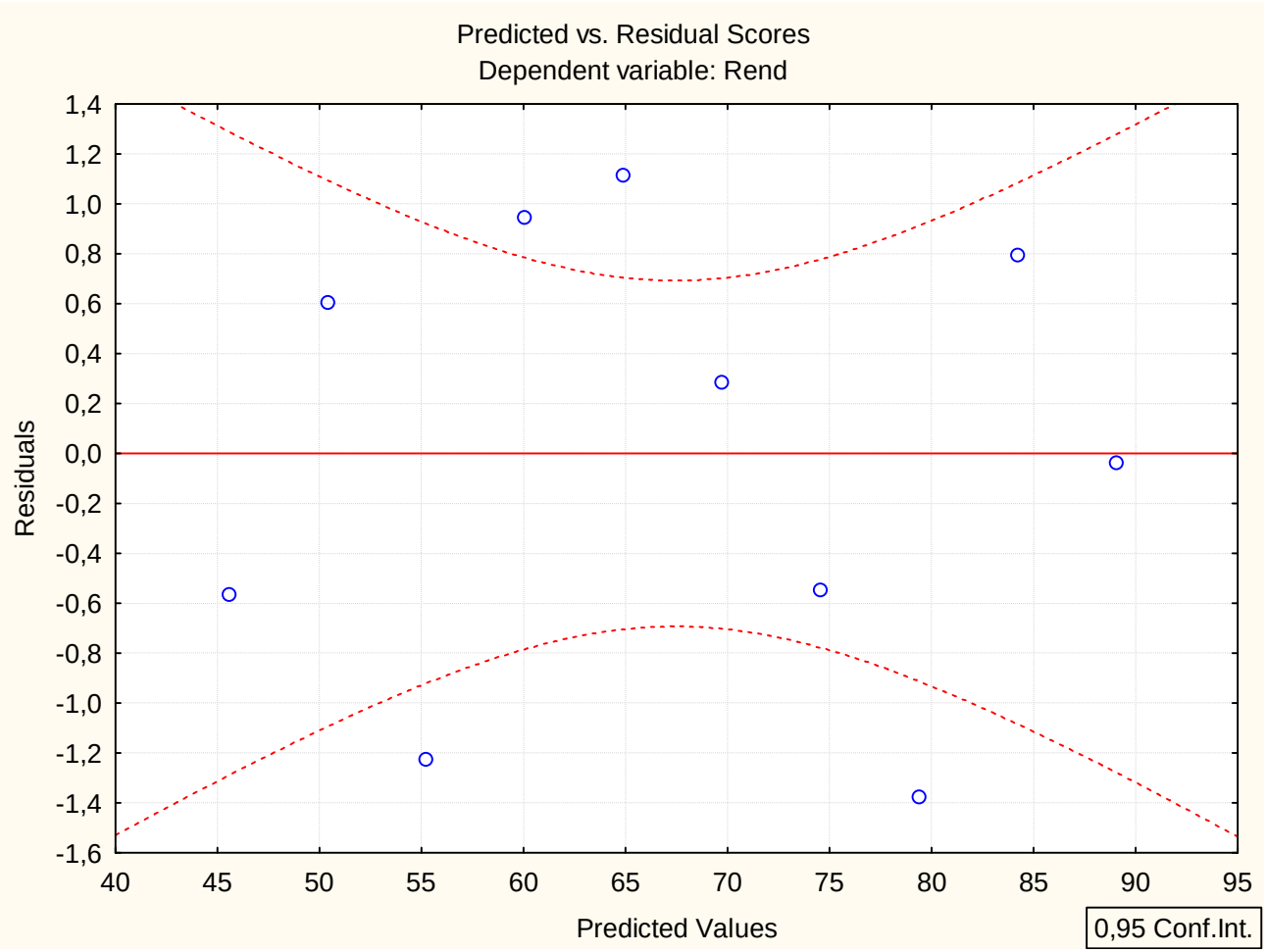
Si les résidus E_i sont normalement distribués alors les erreurs ε_i le sont aussi.

On peut tester si les résidus suivent une loi normale avec :

- Un histogramme.
- Un test de normalité (par ex. Shapiro-Wilk).
- Un graphique de probabilité normal des E_i.

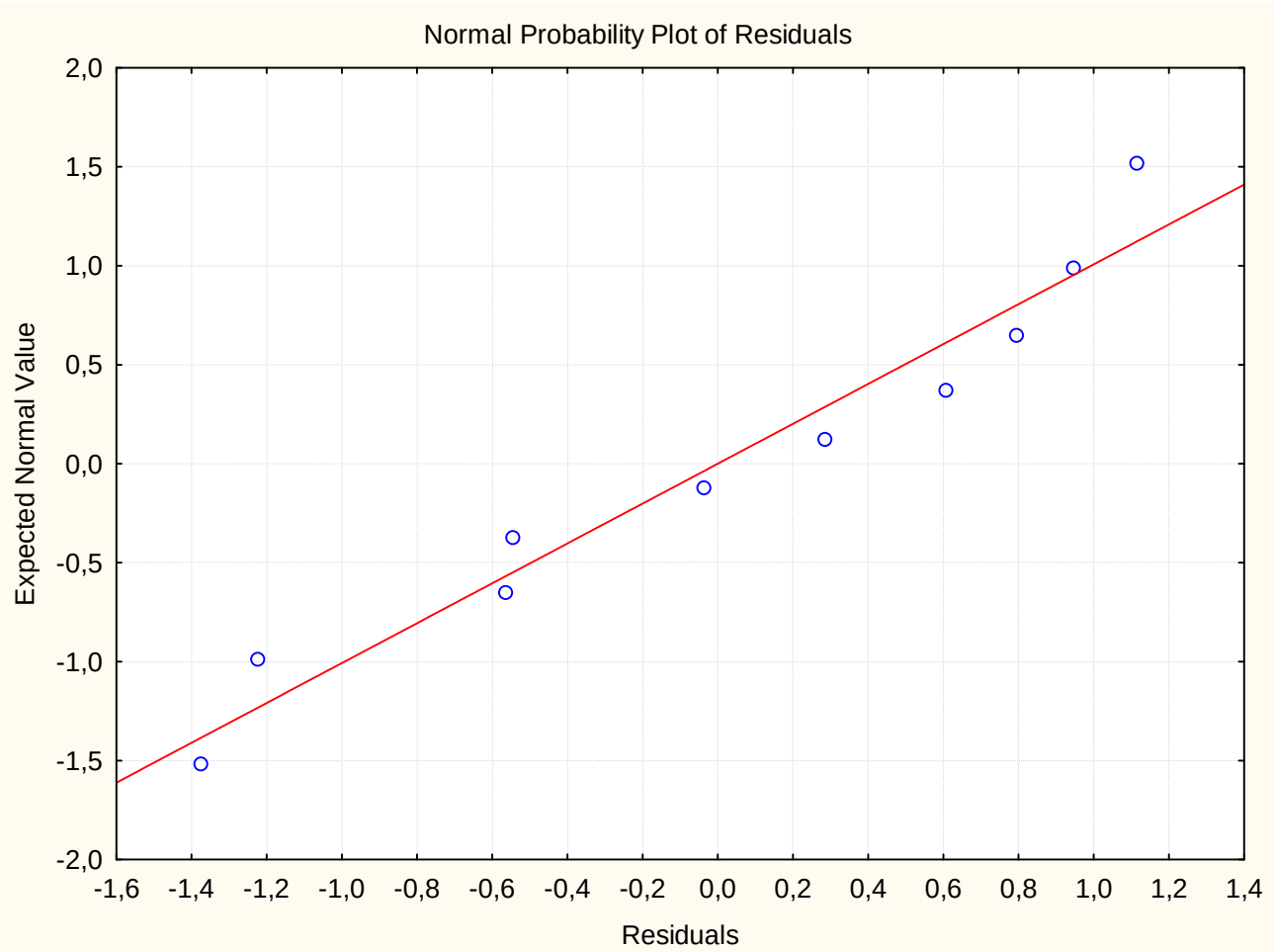
Exemple 1 (suite)

Graphe des points $(\hat{Y}_i, E_i)$:



Exemple 1 (suite)

Graphe de probabilité normale des E_i :



Coefficient de détermination

Le *coefficient de détermination* du modèle de régression linéaire est

$$R^2 = \frac{SS_R}{S_{YY}} = \frac{\hat{\beta}_1^2 S_{XX}}{S_{YY}} = 1 - \frac{SS_E}{S_{YY}} \ .$$

Le coefficient R^2 mesure le pourcentage de la variabilité totale S_{YY} qui est expliquée par le modèle.

Si R^2 est proche de 1, alors le modèle semble adéquat.

Exemple 1 : $R^2 \simeq 99.63\%$.

Coefficient de corrélation

Rappel : La corrélation entre deux variables aléatoires X et Y est mesurée par le coefficient

$$\rho = \frac{\mathsf{Cov}(X, Y)}{\sqrt{\mathsf{V}(X)\mathsf{V}(Y)}} \ .$$

Définition

Le *coefficient de corrélation échantillonnal* est

$$r = \frac{S_{XY}}{\sqrt{S_{XX}S_{YY}}} \ .$$

Le coefficient de corrélation ρ est estimé ponctuellement par r.

Exemple 1 : $r \simeq 99.81\%$.

Interprétation du coefficient de corrélation

On peut montrer que $-1 \leq r \leq 1$.

- Si $r = -1$ ou $r = 1$ alors il y a corrélation parfaite entre X et Y et les points (X_i, Y_i) sont tous sur la droite de régression.
- Si $r = 0$ alors il n'y a pas de corrélation entre X et Y et les points (X_i, Y_i) sont dispersés au hasard.
- Si $0 < r < 1$ alors il y a corrélation positive faible, moyenne ou forte entre X et Y. Dans ce cas, une augmentation de X entraîne une augmentation de Y.
- Si $-1 < r < 0$ alors il y a corrélation négative faible, moyenne ou forte entre X et Y. Dans ce cas, une augmentation de X entraîne une diminution de Y.