R&D

**RECHERCHE ET DEVELOPPEMENT**

**Etude n° 70/00**

# *Modélisation*

**RAPPORT FINAL**

**Chef de projet** : Michèle DESENFANT

Etude réalisée par Caroline Bernard-Michel

**Documents associés** :

**Août 2002**

**R&D - Diffusion générale –**

DRT/SID/Statistiques CBM Régression linéaire

# INTRODUCTION

Dans un laboratoire d'essais, il est très fréquent que l'expérimentateur cherche à ajuster un modèle mathématique à des données expérimentales. En particulier, il cherche souvent à estimer la relation fonctionnelle entre des données d'entrée et des données de sortie. Il existe différentes méthodes pour caractériser une telle relation: interpolations, régression, réseaux de neurones… Dans ce document, nous parlerons uniquement de la régression linéaire. Son avantage par rapport à d'autres méthodes, est qu'elle est basée sur la théorie des probabilités et fournit donc non seulement une courbe ajustant les données, mais aussi des prédictions avec leurs incertitudes. Chaque résultat fourni peut être accompagné de sa variance et de son intervalle de confiance. Mais il faut l'utiliser avec une grande prudence car elle repose sur des hypothèses statistiques qui doivent être absolument vérifiées sous peine d'obtenir des résultats faux. En effet, le calcul des estimateurs, mais surtout de leurs variances et des intervalles de confiance utilise des propriétés qui ne sont valables que si les hypothèses sont respectées. C'est une des raisons principales qui nous a conduit à écrire ce document qui donne les méthodes de vérification des hypothèses. Nous fournissons aussi divers outils statistiques pour détecter les valeurs atypiques ou trop influentes dans la construction du modèle. Ainsi, tous ces éléments permettent à chacun d'apprécier la qualité du modèle choisi.

Dans la première partie, nous étudions la régression linéaire simple avec le détail de tous les calculs, dans la seconde nous présentons la régression linéaire multiple comme généralisation de la régression simple.

Enfin, dans la dernière partie, nous commenterons les outils informatiques disponibles sous Excel.

# 4 Régression linéaire

# REGRESSION LINEAIRE

## INVENTAIRE DES FICHES


**1. INTRODUCTION**

**2. INVENTAIRE DES FICHES**

**3. LA REGRESSION LINEAIRE SIMPLE**

LE MODELE, LES HYPOTHESES, L'OBJECTIF 9

LES MOINDRES CARRES : THEORIE 21

RESIDUS ET INFLUENCES 27

HOMOSCEDASTICITE OU EGALITE DES VARIANCES 39

CORRELATION DES TERMES D'ERREUR 45

NORMALITE DES RESIDUS 47

QUALITE D'AJUSTEMENT, TESTS, INTERVALLES DE CONFIANCE 49

REGRESSION PONDEREE 59

TABLEAU RECAPITULATIF SUR LA REGRESSION LINEAIRE SIMPLE 62

**4. REGRESSION LINEAIRE MULTIPLE** 63

**5. OUTILS INFORMATIQUES** 77

LES OUTILS CLASSIQUES D'EXCEL POUR LA REGRESSION 79

LA REGRESSION PAR LA METHODE DE FORSYTHE (macro Excel) 83

**6. CONCLUSION** 97

**7. ANNEXES** 99

ANNEXE A (les leviers) 101

ANNEXE B (norme NF X 06-050 déc 1995 : étude de la normalité d'une distribution) 103

ANNEXE C (notice droitereg d'excel) 106

ANNEXE D (régression polynomiale) 114

ANNEXE E (exemple : impact d'une valeur atypique) 116

**8. BIBLIOGRAPHIE** 130

# REGRESSION LINEAIRE SIMPLE

# REGRESSION LINEAIRE SIMPLE

## LE MODELE, LES HYPOTHESES, L'OBJECTIF

### 1. Le modèle.

***La droite de régression***

« De nombreuses études consistent à essayer d'expliquer à l'aide d'un modèle, une variable, en fonction d'une autre variable. Dans le cadre de la régression linéaire simple, on écrit alors la relation recherchée sous la forme suivante :

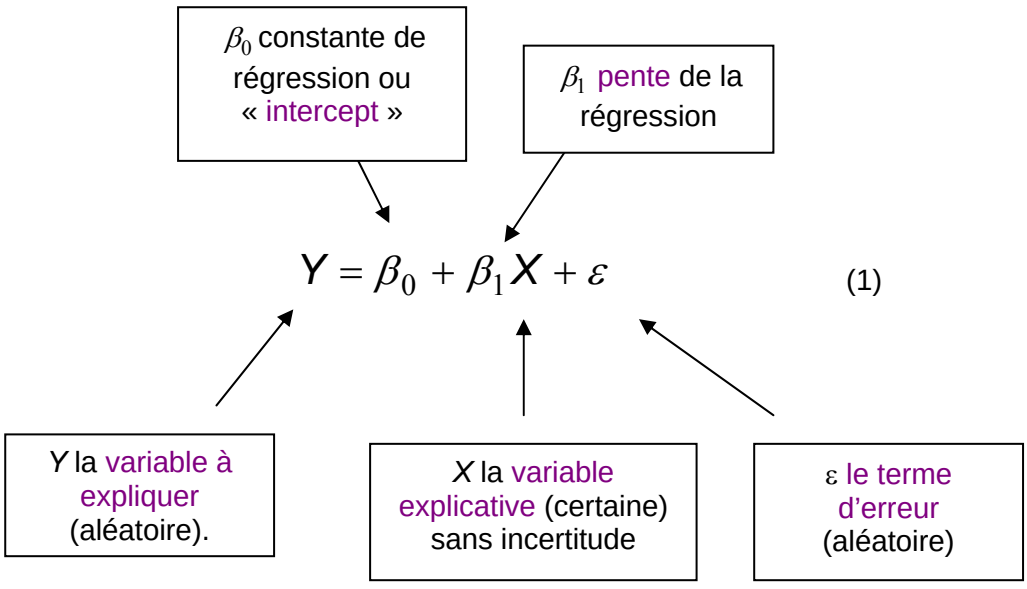
$$Y = \beta_0 + \beta_1 X$$

Exemple : *Y* représente la hauteur des pics mesurée par le chromatographe, et *X* la teneur en éthanol de la substance introduite.

C'est l'équation d'une droite. Elle correspond au modèle « idéal » jamais rencontré en pratique.

Les coefficients $\beta_0$ et $\beta_1$ sont des paramètres inconnus qu'il faut évaluer. Si la relation qui lie *Y* à *X* était parfaitement exacte, il suffirait de connaître les valeurs de *Y* et de *X* pour *2* observations et de résoudre un système d'équations à *2* inconnues pour trouver $\beta_0$ et $\beta_1$.

Cependant, une relation choisie pour expliquer un phénomène donné n'est que rarement exacte. Tout d'abord, un modèle n'est en général qu'une approximation d'un phénomène beaucoup plus complexe. De plus, toute expérience répétée deux fois dans des conditions que l'on croit identiques, ne donne que rarement le même résultat. Les variations sont en général dues à une multitude de facteurs inconnus ou que l'on ne sait pas contrôler. Il est donc nécessaire d'attacher à tout modèle un terme aléatoire qui représente l'écart entre le modèle théorique et la réalité ». Ce terme aléatoire appelé terme d'erreur, est ajouté au modèle comme le montre la relation suivante :



- *X* est la variable dite « explicative » ou « indépendante ». C'est une variable certaine, c'est à dire que les valeurs qu 'elle prend sont fixées par l'expérimentateur et supposées connues **sans incertitude**.

- *Y* est la variable dite « à expliquer » ou « dépendante ». On considère que c'est une variable aléatoire suite au postulat de modélisation sur lequel toute étude statistique est basée :

  « Les données observées sont des réalisations de variables aléatoires » (1)

  On dit que les valeurs de *Y* sont entachées d'une erreur de mesure **aléatoire** notée $\varepsilon$.

## ***Modèle pour un échantillon***

*A la base de toute étude statistique, il y a une population sur laquelle porte l'observation. Lorsqu'il est impossible ou inutile d'étudier l'ensemble de la population, on observe alors une sous-population de taille réduite, en espérant tirer des conclusions généralisables à toute la population. Cette sous-population est appelée échantillon.*
*Pour effectuer une régression linéaire, l'expérimentateur doit disposer d'un échantillon de données ( il est conseillé de faire un plan d'expériences pour obtenir un échantillon tel qu'on ait les meilleures propriétés sur les estimateurs). Ces données sont les observations de la variable expliquée Y pour différentes valeurs fixées de la variable explicative X.*

*Il est important de préciser que le mot échantillon prend deux sens différents, selon qu'on parle des données observées ou du modèle probabiliste. L'hypothèse de modélisation (1) consiste à voir l'échantillon (observé) comme une réalisation d'un échantillon (théorique) d'une certaine loi de probabilité P. En d'autres termes, on considère que les données auraient pu être produites en simulant de façon répétée la loi de probabilité P. Pour éviter les confusions, nous désignerons par données ou échantillon observé, la séquence de nombres recueillie, et échantillon l'échantillon théorique.*

*Si on différencie l'échantillon observé de l'échantillon théorique, c'est parce que leur rôle n'est pas le même : l'échantillon théorique va permettre de trouver des estimateurs et d'utiliser les théorèmes et propriétés statistiques, alors que l'échantillon observé permet de trouver des estimations qui sont des réalisations des estimateurs.*

**Il faut en effet bien distinguer en statistiques l'estimateur qui est une variable aléatoire, de l'estimation qui est la valeur qu'il prend pour un échantillon de données.**

Pour établir les formules des estimateurs et les théorèmes qui leur sont liés, on utilise l'échantillon théorique que l'on note :

### Echantillon (théorique)

⇒ On note $(X_i, Y_i)$ $i \in \{1, n\}$ l'échantillon (théorique) (*n* expériences)

### Termes d'erreur

⇒ On note $\varepsilon_i$ le terme d'erreur associé au couple $(X_i, Y_i)$ $i \in \{1, n\}$.

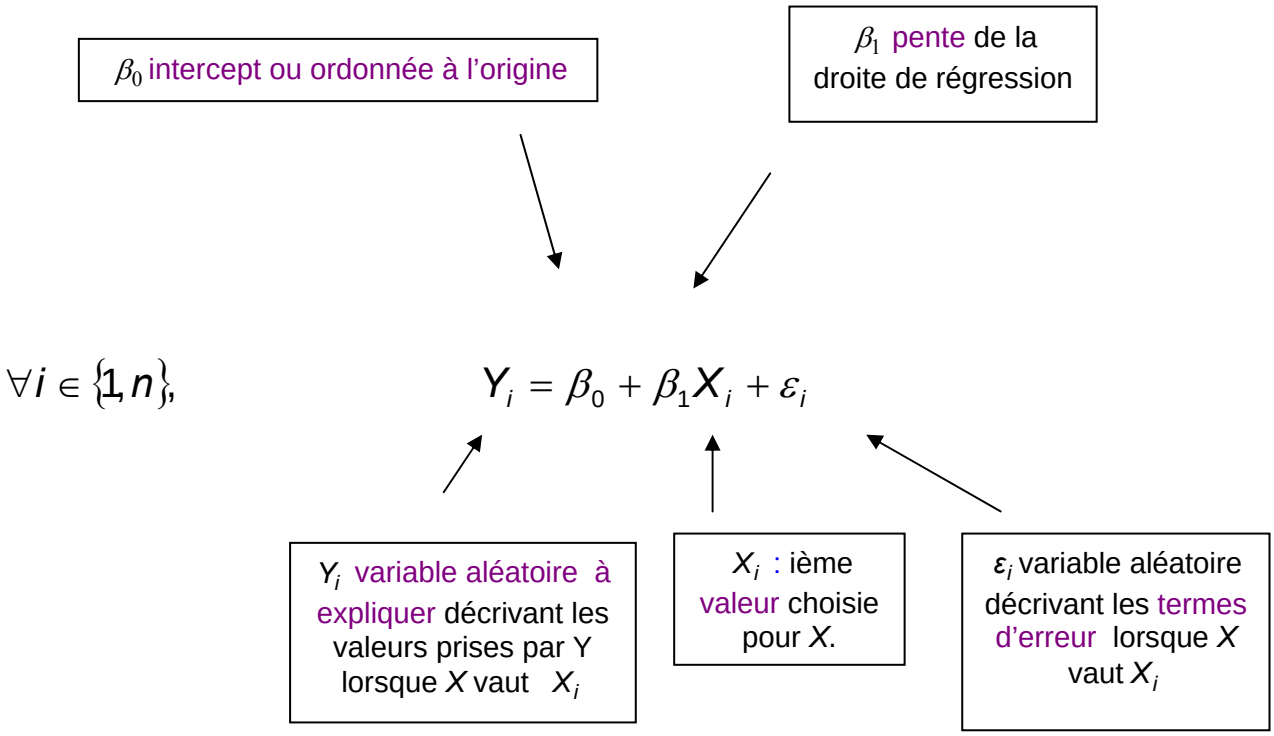
$$\varepsilon_i = Y_i - (\beta_0 + \beta_1 X_i) \quad i \in \{1, n\}$$

**Remarques:**

- $\varepsilon_i$ est une variable aléatoire pour laquelle on n'a aucune observation.

- Lorsqu'on répète une expérience pour une même valeur de $X$ (cas où $X_i = X_j$ pour $i \neq j$), les variables aléatoires $Y_i$ et $Y_j$ suivent la même loi, de même pour $\varepsilon_i$ et $\varepsilon_j$.

**Avec ces notations, et appliqué à un échantillon (théorique) de taille *n* :**
**le modèle de régression linéaire simple devient :**




## *2. Les hypothèses.*

Avant de proposer une méthode pour ajuster une droite à l'échantillon, il est important de préciser les hypothèses sous-jacentes au modèle linéaire simple.

La validité de la plupart des résultats que nous exposons dans la suite dépend directement des hypothèses. Il est donc primordial, avant le choix d'un modèle, de réfléchir à chacune des hypothèses et d'utiliser, après l'estimation des paramètres du modèle, des méthodes adéquates pour les valider.(car il est impossible de valider les hypothèses théoriques, c'est seulement à partir des données qu'on peut les valider)

## *HYPOTHESE 1 :*

- Le modèle doit être « **linéaire par rapport aux paramètres** », c'est à dire que la variable expliquée *Y* s'écrit comme la somme d'une constante ( $\beta_0$ , éventuellement nulle) et d'un paramètre $\beta_1$ multiplié par une fonction de *X* .

**Aucun programme de régression ne pourra fonctionner si cette hypothèse n'est pas vérifiée.**

*Exemples :*

1. $Y = \beta_0 + \beta_1 X^3$ *est un modèle linéaire simple en* $X^3$
2. $Y = \beta_0 + \frac{\beta_1}{X}$ *est un modèle linéaire simple en* $\frac{1}{X}$
3. $Y = \frac{\beta_0}{\beta_1 + X}$ *n'est pas un modèle linéaire*

*Remarques* :

Si on pose $X^* = X^3$ pour l'exemple 1 et $X^* = \frac{1}{X}$ pour l'exemple 2, on retrouve bien le modèle de régression simple $Y = \beta_0 + \beta_1 X^*$.

## HYPOTHESE 2 :

Le modèle doit être bien spécifié, c'est à dire qu'il ne faut pas avoir oublié des variables explicatives. ( voir régression multiple si nécessaire)

## HYPOTHESE 3 :

- La variable explicative ***X* est déterministe**, c'est à dire que les valeurs prises par *X* sont parfaitement connues (contrairement à une variable aléatoire). On parle alors de modèle à effets fixes ou encore de modèle à facteur contrôlé. Il existe aussi des modèles à effets aléatoires pour lesquels les valeurs de *X* sont aléatoires (Pour ces modèles, une fiche sera rédigée par la suite, pour le moment se renseigner au service statistiques)

## HYPOTHESE 4 :

- **L'espérance des erreurs** associées à chaque couple d'observations est **nulle** :

$$E(\varepsilon_i) = 0 \qquad \forall i \in \{1, n\}$$

Cela implique que la moyenne de $Y_i$ lorsque $X = X_i$ est $\beta_o + \beta_1 X_i$ . Plus généralement, on a :

$$\boxed{E(Y) = \beta_0 + \beta_1 X}$$

On verra qu'avec cette hypothèse, l'estimateur des moindres carrés est **sans biais.** (une fiche sur les propriétés des estimateurs sera ajoutée par la suite).

## HYPOTHESE 5 :

- La **variance des erreurs** associées à chaque couple d'observations est **constante**. On a donc :

$$Var(\varepsilon_i) = \text{ constante } = \sigma^2 \qquad \forall i \in \{1,n\}$$

L'écart-type (=$\sqrt{Var(\varepsilon_i)}$) des termes d'erreur est donc constant.

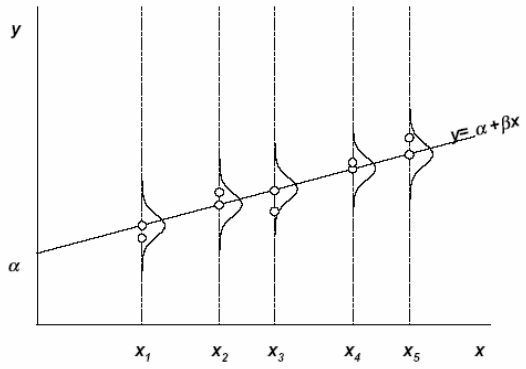
Lorsque cette hypothèse est vérifiée, on dit qu'il y a « **homoscédasticité** » (homogénéité des variances)

Pour tester l'homoscédasticité, se reporter à la fiche « Homoscédasticité » ou égalité des variances ».

Exemples :

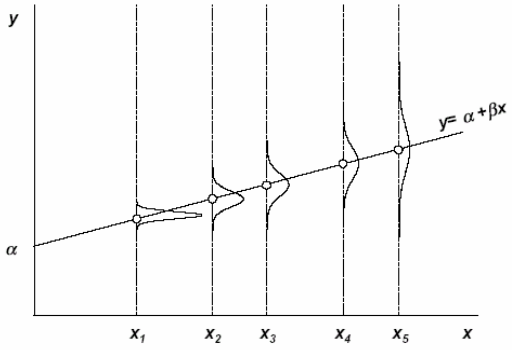
Dans le cas où la variable expliquée suit pour chaque niveau une loi normale.

1. les variances du terme d'erreur sont égales pour chaque niveau




**Homoscédasticité**

2. les variances du terme d'erreur ne sont pas égales




**Hétéroscédasticité**

**Si l'homoscédasticité n'est pas vérifiée, se reporter à la fiche "**REGRESSION PONDEREE".

**HYPOTHESE 6 :**

- Les termes d'erreur ne sont pas corrélés. Il ne doit y avoir aucun lien entre l'erreur commise sur une mesure et les mesures effectuées précédemment. Mathématiquement, cela se traduit par :

$$Cov(\varepsilon_i, \varepsilon_j) = 0 \quad pour\ i \neq j$$

Pour tester cette hypothèse, se reporter à la fiche « corrélation des termes d'erreur »

**HYPOTHESE 7 ( FACULTATIVE ) :**

- Les variables aléatoires $\varepsilon_i$ suivent une loi normale $N(0,\sigma^2)$.
  Cette hypothèse est nécessaire pour l'utilisation des tests statistiques (Student, Fisher …) et la construction d'intervalles de confiance.

Remarque : les variables $Y_i$ suivent alors toutes une loi normale
Pour tester cette hypothèse se reporter à la fiche « Normalité des résidus »

## 3. Les objectifs.

### Les estimateurs

L' objectif de la régression est d'estimer les coefficients du modèle ($\beta_0$ et $\beta_1$) ainsi que la variance $\sigma^2$ de l'erreur $\varepsilon$ à partir d'un échantillon.

$\Rightarrow$ on appelle $b_0$ l'estimateur de $\beta_0$

$\Rightarrow$ on appelle $b_1$ l'estimateur de $\beta_1$

$\Rightarrow$ on appelle $s^2$ l'estimateur de $\sigma^2$

Ainsi, la droite de régression « réelle » $Y = \beta_0 + \beta_1 X$, inconnue, est estimée par la droite :

$$\hat{Y} = b_0 + b_1 X$$

Dans ce document, nous utilisons toujours l'échantillon théorique afin d'obtenir les estimateurs et leurs propriétés. Mais de façon concrète, l'expérimentateur a besoin des estimations. Pour les obtenir, il lui suffit, dans les formules des estimateurs, de remplacer les variables par les données qu'il a obtenues. En toute rigueur, il faudrait différencier les notations utilisées pour les estimateurs et celles pour les estimations, de même il faudrait différencier les notations utilisées pour les données et celles pour les variables aléatoires dont elles sont les réalisations, mais cela complique et alourdit le document, c'est pourquoi nous garderons les mêmes notations, comme dans la plupart des ouvrages sur la régression.

Remarques :

- $\hat{Y} = b_0 + b_1 X$ est appelée « droite d'estimation » ou encore « droite de régression de $Y$ en $X$ ».
- différents jeux de données conduisent à des estimations différentes de $\beta_0$ et $\beta_1$. Les estimations sont cependant toutes très proches si on a un assez grand nombre de données.

Pour un échantillon, $\hat{Y}_i = b_0 + b_1 X_i$ est l'estimateur de la valeur moyenne prise par la variable $Y$ quand la variable $X$ prend la valeur $X_i$.

Il existe différentes méthodes pour calculer les estimateurs des coefficients. Nous exposons ici celle des moindres carrés.

Elle consiste à minimiser la sommes des carrés des écarts $Y_i - \hat{Y}_i$ , c'est à dire qu'on cherche :

$$\underset{b_0, b_1}{Min} \sum_{i=1}^{n} (Y_i - b_0 - b_1 X_i)^2$$

Ces écarts $Y_i - \hat{Y}_i$ sont appelés les résidus, ce sont des variables aléatoires que nous noterons $e_i$. Ils ne sont pas égaux aux termes d'erreur et contrairement à ces derniers, on peut connaître leur valeur pour chaque échantillon de données.

$\Rightarrow$ On note $e_i = Y_i - \hat{Y}_i$ les résidus

**Remarque:** on trouve souvent dans les ouvrages de statistiques que la méthode des moindres carrés consiste à minimiser la somme des carrés des erreurs $\varepsilon_i$ :

$$\underset{b_0, b_1}{Min} \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2$$

Cela conduit exactement aux mêmes estimateurs.

Soient les notations suivantes :

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i \qquad \bar{Y} = \frac{1}{n} \sum_{i=1}^{n} Y_i$$

$$s_x^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2 \qquad s_Y^2 = \frac{1}{n-1} \sum_{i=1}^{n} (Y_i - \bar{Y})^2$$

$$s_{xy}^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})$$

La méthode des moindres carrés conduit aux estimateurs suivants :

| Paramètres de la droite théorique | Estimateur | Variance de l'estimateur |
|---|---|---|
| $\beta_0$ | $b_0 = \overline{Y} - b_1\overline{X}$ | $Var(b_0) = s^2\left(\frac{1}{n} + \frac{\overline{X}^2}{\sum_{i=1}^{n}(X_i - \overline{X})^2}\right)$ |
| $\beta_1$ | $b_1 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^{n}(X_i - \overline{X})} = \frac{s_{xy}}{s_x^2}$ | $Var(b_1) = \frac{s^2}{\sum_{i=1}^{n}(X_i - \overline{X})^2}$ |
| | | $Cov(b_0, b_1) = s^2 \frac{-\overline{X}}{\sum_{i=1}^{n}(X_i - \overline{X})^2}$ |
| $\sigma^2$ | $s^2 = \frac{1}{n-2}\sum_{i=1}^{n}(Y_i - b_0 - b_1X_i)^2$ $= \frac{1}{n-2}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 = \frac{1}{n-2}\sum_{i=1}^{n}e_i^2$ | |

**Remarques :**

- Plus la variance du terme d'erreur du modèle ($s^2$, appelée variance résiduelle) est faible, meilleurs seront les estimateurs.
- Les variances et les covariances diminuent quand la taille de l'échantillon augmente. Un grand échantillon donne donc de meilleures estimations.
- Les variances et les covariances sont inversement proportionnelles à la dispersion des *X* autour de leur moyenne. L'expérimentateur aura donc intérêt à placer ses essais aux extrémités de ce domaine pour obtenir des estimations précises.

## *Les prédictions.*

On utilise souvent les modèles linéaires pour faire de la prédiction, c'est à dire calculer la valeur de *Y* lorsqu'on a choisi une valeur pour *X*.
Dans ces fiches, nous utilisons le terme de prédiction (prediction en anglais), mais il est important de signaler que la littérature utilise aussi pour les mêmes résultats le mot prévision (forecasting). La différence entre les deux mots n'est pas clairement expliquée dans les ouvrages et donne lieux à de nombreuses discussions.

Dans le modèle de régression linéaire simple, il faut distinguer deux sortes de prédictions :

- $\hat{Y}_k$ la prédiction sur la moyenne de $Y$ pour X= $X_k$ . C'est l'estimateur de l'espérance de $Y$ sachant que X= $X_k$ (notée mathématiquement $E(Y/X=X_k)$ ), c'est à dire l'estimateur de la valeur moyenne prise par la variable $Y$ quand la variable $X$ prend la valeur $X_k$ .

- $\hat{\hat{Y}}_k$ la prédiction sur une nouvelle valeur de $Y$ pour X= $X_k$ .(résultat d'un nouvel essai au point X= $X_k$ ). C'est l'estimateur de $Y$ sachant que X= $X_k$ (notée mathématiquement $Y/X=X_k$ ), c'est à dire l'estimateur de la valeur que prendra la variable $Y$ pour X= $X_k$ .

Les estimateurs $\hat{Y}_k$ et $\hat{\hat{Y}}_k$ sont égaux, ce sont des points qui se situent sur la droite de régression, mais ils n'ont pas la même variance (c'est pourquoi il est important de bien différencier les deux).

| | **ESTIMATEUR** | **VARIANCE DE L'ESTIMATEUR** |
|---|---|---|
| $E(Y/X=X_k)$ $(=\beta_0+\beta_1 X_k)$ | $\hat{Y}_k = b_0 + b_1 X_k$ | $Var(\hat{Y}_k) = s^2\left(\frac{1}{n}+\frac{(X_k-\overline{X})^2}{\sum_{i=1}^{n}(X_i-\overline{X})^2}\right)$ |
| $Y/(X=X_k)$ $(=\beta_0+\beta_1 X_k+\varepsilon_k)$ | $\hat{\hat{Y}}_k = b_0 + b_1 X_k$ +e avec e estimateur de $\varepsilon_k$ (e=0, var(e)= $\sigma^2$ ) | $Var(\hat{\hat{Y}}_k) = s^2\left(1+\frac{1}{n}+\frac{(X_k-\overline{X})^2}{\sum_{i=1}^{n}(X_i-\overline{X})^2}\right)$ |

Pour connaître les intervalles de confiance liés à ces prédictions, se reporter à la fiche : "QUALITE DE L'AJUSTEMENT"

***Résumé notations :***

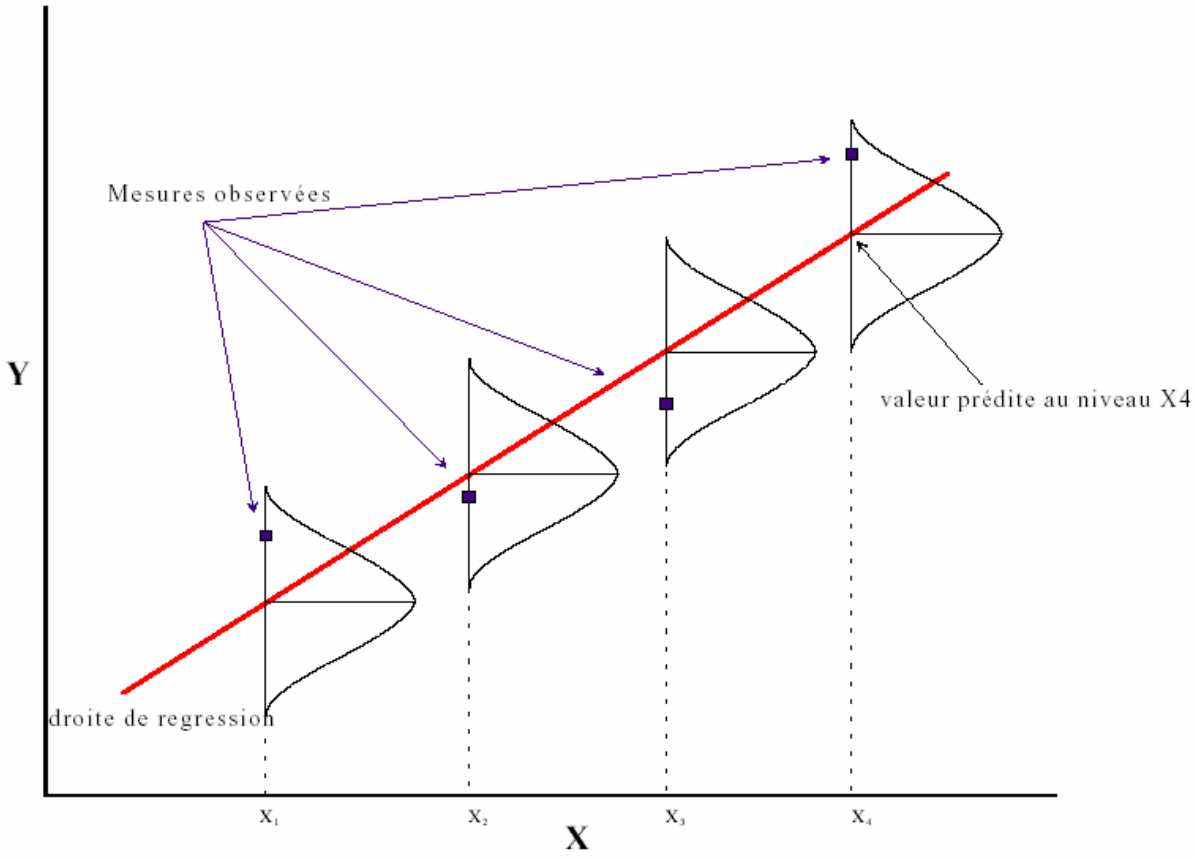
Enfin, pour résumer les notations qui sont utilisées tout au long de ce document, voici un tableau des notations et un schéma sur la régression.

## TABLEAU RECAPITULATIF DES NOTATIONS

| Notations | Variable | constante | aléatoire | certaine ou déterministe | observable (O) ou calculable (C) |
|---|---|---|---|---|---|
| $Y$ (population) | × | | × | | O |
| $Y_i$ variable à expliquer (échantillon) | × | | × | | O |
| $X$ (population) variable explicative | × | | | × | O |
| $X_i$ variable explicative (échantillon) | | | | | O |
| $\varepsilon_i = Y_i - \beta_0 - \beta_1 X_i$ variable d'erreur | × | | × | | |
| $\beta_0$ intercept du modèle | | × | | × | |
| $\beta_1$ pente du modèle | | × | | × | |
| $e_i = Y_i - b_0 - b_1 X_i$ résidus | × | | × | | C |
| $b_0$ estimateur de l'intercept | × | | × | | C |
| $b_1$ estimateur de la pente | × | | × | | C |
| $s^2$ estimateur de la variance d'erreur | × | | × | | C |

# SCHEMA DU MODELE DE REGRESSION SIMPLE

Si on suppose par exemple que tous les $\varepsilon_i$ suivent une loi normale centrée $N(0,\sigma^2)$, le modèle de régression linéaire implique que les $Y_i$ suivent aussi une loi normale de moyenne $\beta_o + \beta_1 X_i$. On a donc le schéma suivant :



# REGRESSION LINEAIRE SIMPLE

## LES MOINDRES CARRES : THEORIE

| **Note technique n° 004** | |
|---|---|
| **Objet : Formulaire pour les régressions linéaires** | |
| Date de diffusion : **1er mai 2002** | Nombre de pages : **6** |
| Annule et remplace : *I* | Rédaction : **Ch. Ranson** |
| Destinataires : **diffusion générale** | **Visa :** |

Dans l'exemple **H.3**, le GUM, *Guide pour l'expression de l'incertitude de mesure* (ISO, première édition, 1995), utilise une série de formules (H.13a à H.13g). Toutes ces formules, application de la méthode des moindres carrés à la régression linéaire, sont démontrées ci-dessous.

### 1. Notations

Soient $n$ paires d'observations $X_i \quad Y_i$

la moyenne des $X_i$ est notée $$\overline{X} = \frac{\sum X_i}{n}$$

de même, la moyenne des $Y_i$ est notée $$\overline{Y} = \frac{\sum Y_i}{n}$$

On note également

$$S_{XY} = \sum (X_i - \overline{X})(Y_i - \overline{Y}) = \sum (X_i - \overline{X})Y_i = \sum X_i (Y_i - \overline{Y})$$
$$= \sum X_i Y_i - \overline{X} \sum Y_i - \overline{Y} \sum X_i + n\overline{X}\overline{Y}$$
$$= \sum X_i Y_i - n\overline{X}\overline{Y} = \sum X_i Y_i - \frac{(\sum X_i)(\sum Y_i)}{n}$$

de même

$$S_{XX} = \sum (X_i - \overline{X})^2 = \sum (X_i - \overline{X})X_i = \sum X_i^2 - n\overline{X}^2 = \sum X_i^2 - \frac{(\sum X_i)^2}{n}$$

**2. Rappel sur la covariance de résultats simultanés**

Si les deux quantités *a* et *c* se déduisent des variables communes $Y_1 \ldots Y_i \ldots Y_n$ par les fonctions :

$$\begin{aligned} a &= a_1 Y_1 + \ \ldots \ + a_i Y_i + \ \ldots \ + a_n Y_n \\ c &= c_1 Y_1 + \ \ldots \ + c_i Y_i + \ \ldots \ + c_n Y_n \end{aligned}$$

avec $V(Y_i) = \sigma^2$ quel que soit *i*

et $Y_i$ non corrélé avec $Y_j$ pour tout $i \neq j$

alors $\text{cov}(a,c) = (a_1 c_1 + \ldots + a_i c_i + \ldots + a_n c_n)\sigma^2$

ce qui peut aussi s'écrire $\text{cov}(a,c) = \sigma^2 \sum a_i c_i$

**3. La régression linéaire par moindres carrés**

La droite "idéale" serait définie par les paramètres inconnus $\beta_0$ et $\beta_1$ ayant ainsi pour équation :

$$Y = \beta_0 + \beta_1 X \qquad (1)$$

On va chercher des estimations de $\beta_0$ et $\beta_1$, qui seront notées $b_0$ et $b_1$. Ces valeurs seront utilisées pour calculer les valeurs prédites de *Y* notées $\hat{Y}$ :

$$\hat{Y} = b_0 + b_1 X \qquad (2)$$

Pour chacune des *n* paires d'observations $(X_i, Y_i)$ on peut écrire :

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \qquad (3)$$

$\varepsilon_i$ représentant l'écart entre la valeur observée $Y_i$ et l'ordonnée de la droite "idéale".

Considérons la somme des carrés des déviations par rapport à cette droite :

$$S = \sum \varepsilon_i^2 = \sum (Y_i - \beta_0 - \beta_1 X_i)^2 \qquad (4)$$

on adopte pour valeurs de $\beta_0$ et $\beta_1$ les valeurs de $b_0$ et $b_1$ telles que la somme *S* soit minimale (de là vient la dénomination "moindres carrés").

**4. Calcul des coefficients de régression**

Cette somme *S* dont on recherche le minimum étant une fonction de $\beta_0$ et $\beta_1$, on calcule sa dérivée par rapport à chacune de ces deux variables :

$$\begin{cases} \dfrac{\partial S}{\partial \beta_0} = -2\sum\left(Y_i - \beta_0 - \beta_1 X_i\right) \\ \dfrac{\partial S}{\partial \beta_1} = -2\sum X_i\left(Y_1 - \beta_0 - \beta_1 X_i\right) \end{cases} \qquad (5)$$

et les racines des "équations normales" fournissent ainsi les estimations de $b_0$ et $b_1$ :

$$\begin{cases} \sum\left(Y_i - b_0 - b_1 X_i\right) = 0 \\ \sum\left(Y_i - b_0 - b_1 X_i\right) X_i = 0 \end{cases} \qquad (6)$$

$$\begin{cases} \sum Y_i - n\, b_0 - b_1 \sum X_i = 0 \\ \sum X_i Y_i - b_0 \sum X_1 - b_1 \sum X_i^2 = 0 \end{cases} \qquad (7)$$

$$\begin{cases} b_0\, n + b_1 \sum X_i = \sum Y_1 & (8a) \\ b_0 \sum X_i + b_1 \sum X_i^2 = \sum X_i Y_i & (8b) \end{cases}$$

De (8a) on déduit $$b_0 = \frac{1}{n}\left(\sum Y_i - b_1 \sum X_i\right) \qquad (9a)$$

ou $$b_0 = \overline{Y} - b_1 \overline{X} \qquad (9b)$$

De (8b) et de (9a) on déduit :

$$\frac{1}{n}\left(\sum X_i\right)\left(\sum Y_i\right) - \frac{1}{n} b_1 \left(\sum X_i\right)^2 + b_1 \sum X_i^2 = \sum X_i Y_i$$

soit (H.13b) $$b_1 = \frac{\sum X_i Y_i - \frac{1}{n}\left(\sum X_i\right)\left(\sum Y_i\right)}{\sum X_i^2 - \frac{1}{n}\left(\sum X_i\right)^2} \qquad (10a)$$

ou encore $$b_1 = \frac{\sum\left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)}{\sum\left(X_i - \overline{X}\right)^2} = \frac{S_{XY}}{S_{XX}} \qquad (10b)$$

De (9a) et de (10a) on déduit (H.13a) :

$$b_0 = \frac{\sum Y_i}{n} - \frac{\sum X_i}{n}\left[\frac{n\sum X_i Y_i - \left(\sum X_i\right)\left(\sum Y_i\right)}{n\sum X_i^2 - \left(\sum X_i\right)^2}\right] = \frac{\left(\sum X_i^2\right)\left(\sum Y_i\right) - \left(\sum X_i\right)\left(\sum X_i Y_i\right)}{n\sum X_i^2 - \left(\sum X_i\right)^2} \qquad (9c)$$

De (2) et de (9b) on déduit : $$\hat{Y} = \overline{Y} + b_1\left(X - \overline{X}\right) \qquad (11)$$

**5. Variance de $b_1$**

Reprenant l'expression de $b_1$ démontrée en (10b) :

$$b_1 = \frac{\sum(X_i - \overline{X})(Y_i - \overline{Y})}{\sum(X_i - \overline{X})^2} = \frac{\sum(X_i - \overline{X})Y_i}{\sum(X_i - \overline{X})^2} = \frac{(X_1 - \overline{X})}{\sum(X_i - \overline{X})^2} Y_1 + \ \dots \ + \frac{(X_n - \overline{X})}{\sum(X_i - \overline{X})^2} Y_n$$

si l'on accepte les hypothèses suivantes :

dans $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$

1°) $\varepsilon_i$ est une variable aléatoire telle que $E(\varepsilon_i) = 0$ et $V(\varepsilon_i) = \sigma^2$

2°) $\varepsilon_i$ et $\varepsilon_j$ sont non corrélées, ou $\text{cov}(\varepsilon_i, \varepsilon_j) = 0$ quels que soient *i* et *j*

alors on en déduit :

1°) $E(Y_i) = \beta_0 + \beta_1 X_i$

2°) $V(Y_i) = \sigma^2$ quel que soit *i*

3°) $Y_i$ et $Y_j$ sont non corrélés quels que soient *i* et *j*

D'où (H.13d) :

$$V(b_1) = \left[ \sum \left( \frac{X_i - \overline{X}}{\sum(X_i - \overline{X})^2} \right)^2 \right] \sigma^2 = \frac{\sigma^2}{\sum(X_i - \overline{X})^2} \qquad (12)$$

**6. Variance de $b_0$**

Reprenant l'expression (9b) de $b_0$ on écrit sa variance :

$$V(b_0) = V(\overline{Y} - b_1 \overline{X}) = V(\overline{Y}) + \overline{X}^2 V(b_1) - 2\overline{X}\,\text{cov}(\overline{Y}, b_1)$$

Mais $\text{cov}(\overline{Y}, b_1) = 0$ . En effet

$$\overline{Y} = \frac{1}{n} Y_1 + \quad \dots \quad + \frac{1}{n} Y_n$$

et

$$b_1 = \frac{(X_1 - \overline{X})}{\sum(X_i - \overline{X})^2} Y_1 + \quad \dots \quad + \frac{(X_n - \overline{X})}{\sum(X_i - \overline{X})^2} Y_n$$

et comme les $Y_i$ et $Y_j$ ne sont pas corrélés, et que $V(Y_i) = \sigma^2$ quel que soit *i*, alors (d'une part voir le rappel au paragraphe 2, d'autre part se souvenir de ce que $\sum(X_i - \overline{X}) \equiv 0$ ) :

$$\text{cov}(\overline{Y}, b_1) = \sigma^2 \sum \left[ \frac{1}{n} \frac{(X_i - \overline{X})}{\sum(X_i - \overline{X})^2} \right] = 0$$

Donc, utilisant le résultat (12),

$$V(b_0) = V(\overline{Y}) + \overline{X}^2 V(b_1) = \frac{\sigma^2}{n} + \overline{X}^2 \frac{\sigma^2}{\sum (X_i - \overline{X})^2}$$

$$V(b_0) = \sigma^2 \left( \frac{1}{n} + \frac{\overline{X}^2}{\sum (X_i - \overline{X})^2} \right) \qquad (13)$$

ou encore, selon les notations du *Guide...* (H.13c)

$$V(b_0) = \frac{\sigma^2 \left[ \sum (X_i - \overline{X})^2 + n\overline{X}^2 \right]}{n \sum (X_i - \overline{X})^2} = \frac{\sigma^2 \sum X_i^2}{n \sum (X_i - \overline{X})^2} \qquad (14)$$

## 7. Covariance et coefficient de corrélation de $b_0$ et $b_1$

Utilisant (9b), le rappel du paragraphe 2 et (12)

$$\mathrm{cov}(b_0, b_1) = \mathrm{cov}(\overline{Y} - b_1 \overline{X}\ , b_1) = -\overline{X}\, V(b_1) = \frac{-\overline{X}\sigma^2}{\sum (X_i - \overline{X})^2} \qquad (15)$$

On en déduit, avec (14) et (12), le coefficient de corrélation

$$r(b_0, b_1) = \frac{\mathrm{cov}(b_0, b_1)}{s(b_0)\, s(b_1)} = \frac{-\overline{X}\sigma^2}{\sum (X_i - \overline{X})^2 \sqrt{\dfrac{\sigma^2 \sum X_i^2}{n \sum (X_i - \overline{X})^2} \dfrac{\sigma^2}{\sum (X_i - \overline{X})^2}}}$$

ou, après simplification

$$r(b_0, b_1) = \frac{-\overline{X}}{\sqrt{\dfrac{\sum X_i^2}{n}}}$$

c'est-à-dire (H.13e) $$r(b_0, b_1) = \frac{-\sum X_i}{\sqrt{n \sum X_i^2}}$$

## 8. Variance de la valeur estimée

Reprenant la formule (2)

$$V(\hat{Y}) = V(b_0 + b_1 X) = V(b_0) + X^2 V(b_1) + 2X\, \mathrm{cov}(b_0, b_1)$$

et utilisant (13), (12) et (15)

$$V(\hat{Y}) = \sigma^2 \left( \frac{1}{n} + \frac{\overline{X}^2}{\sum (X_i - \overline{X})^2} \right) + \frac{X^2 \sigma^2}{\sum (X_i - \overline{X})^2} - 2 \frac{X \overline{X} \sigma^2}{\sum (X_i - \overline{X})^2}$$

donc $$V(\hat{Y}) = \sigma^2 \left[ \frac{1}{n} + \frac{(X - \overline{X})^2}{\sum (X_i - \overline{X})^2} \right] \qquad (16)$$

**9. Estimation de la variance des résidus**

Il s'agit d'estimer $\sigma^2 = V(\varepsilon_i)$, variance des écarts des observations à la droite "idéale".

On dispose des écarts à la droite estimée $Y_i - \hat{Y}_i$. On estimera $\sigma^2$ au moyen du quotient de la somme des carrés de ces écarts par le nombre de degrés de liberté correspondant à cette somme.

Or, de l'égalité

$$(Y_i - \overline{Y}) = (Y_i - \hat{Y}_i) + (\hat{Y}_i - \overline{Y})$$

on déduit (élévation au carré et sommation pour tous les points)

$$\begin{aligned} \sum (Y_i - \overline{Y})^2 &= \sum \left[ (Y_i - \hat{Y}_i) + (\hat{Y}_i - \overline{Y}) \right]^2 \\ &= \sum (Y_i - \hat{Y}_i)^2 + 2\sum (Y_i - \hat{Y}_i)(\hat{Y}_i - \overline{Y}) + \sum (\hat{Y}_i - \overline{Y})^2 \end{aligned}$$

de (2) et (9b) on déduit $\hat{Y} = \overline{Y} + b_1(X - \overline{X})$

et donc, d'une part $\sum (\hat{Y}_i - \overline{Y})^2 = b_1^2 \sum (X_i - \overline{X})^2 = b_1^2 S_{XX}$

et, d'autre part

$$\begin{aligned} 2\sum (Y_i - \hat{Y}_i)(\hat{Y}_i - \overline{Y}) &= 2b_1 \sum \left[ (Y_i - \overline{Y}) - b_1 (X_i - \overline{X}) \right] (X_i - \overline{X}) \\ &= 2b_1 (S_{XY} - b_1 S_{XX}) \\ &= 0 \end{aligned}$$

d'après (10b).

Finalement :

$$\sum (Y_i - \overline{Y})^2 = \sum (Y_i - \hat{Y}_i)^2 + \sum (\hat{Y}_i - \overline{Y})^2$$

$\sum (Y_i - \overline{Y})^2$ est la somme des carrés des écarts des valeurs observées par rapport à la moyenne, elle a $n - 1$ degrés de liberté.

$\sum (\hat{Y}_i - \overline{Y})^2$ est la somme des carrés des écarts des valeurs prédites par rapport à la moyenne ; c'est la somme des carrés due à (expliquée par) la régression ; comme elle vaut $b_1^2 S_{XX}$ elle n'a qu'un seul degré de liberté.

$\sum (Y_i - \hat{Y}_i)^2$ qui est la somme des carrés des résidus (part de la somme des carrés initiale inexpliquée par la régression) a en conséquence $n - 2$ degrés de liberté.

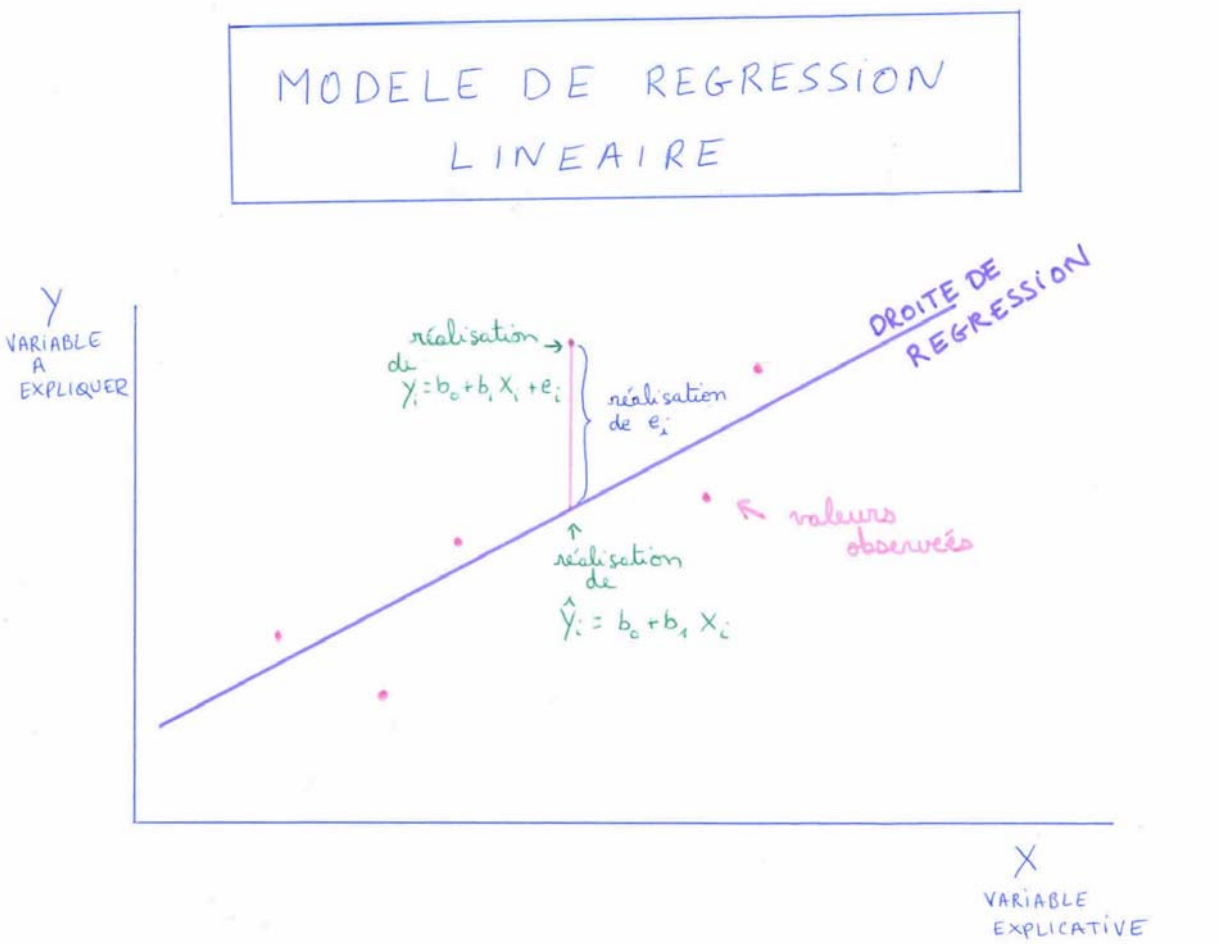
$\sigma^2 = V(\varepsilon_i)$ sera donc estimé (voir H.13f) par

$$s^2 = \frac{\sum (Y_i - \hat{Y}_i)^2}{n - 2}$$

# REGRESSION LINEAIRE SIMPLE

## RESIDUS ET INFLUENCES

Précédemment, on a noté $\varepsilon_i = Y_i - \beta_0 - \beta_1 X_i \, , i \in \{1,n\}$ les termes d'erreur aléatoire d'une régression . Ces termes permettent de détecter des défaillances dans le modèle mais ils ne peuvent pas être calculés puisqu'ils dépendent de $\beta_0$ et $\beta_1$ inconnus. C'est pourquoi on introduit les résidus $e_i = Y_i - b_0 - b_1 X_i \, , i \in \{1,n\}$ qui dépendent de l'échantillon. Ce sont des estimateurs convergents ( on dit aussi consistants) des erreurs $\varepsilon_i$ .De plus, on a accès à des réalisations des $e_i$ en utilisant l'échantillon de données.



On ne peut déterminer la valeur des résidus qu'après avoir effectué une régression, mais on les utilise souvent pour vérifier les hypothèses de la régression.
Ils permettent par exemple de vérifier :

- l'homoscédasticité (voir fiche « Homoscédasicité ou égalité des variances »)
- la corrélation (voir fiche « corrélation »)
- la normalité des termes d'erreur (voir fiche « normalité des erreurs »).
- Ils mettent aussi en évidence la présence d'observations atypiques.

Il existe différents types de résidus dont les formules sont données dans le « tableau des résidus ».(quelques pages plus loin). Ils nécessitent des notations introduites dans l'annexe A.

*En régression linéaire, il existe des méthodes qui permettent de détecter les valeurs atypiques (« outliers ») , c'est à dire les données qui s'éloignent fortement de la droite de régression, et les valeurs influentes, c'est à dire celles qui apportent une forte contribution dans le calcul de la droite de régression. Ces méthodes sont souvent réunies sous le nom de « diagnostics »*

## ***Les valeurs atypiques***

Remarque : Les graphes qui suivent ont tous été réalisés avec le même échantillon de données.

Notations : Le calcul des résidus utilise les notations suivantes :

- $h_{ii}$ les leviers (voir annexe A) $\left(h_{ii} = \frac{1}{n} + \frac{(X_i - \overline{X})^2}{\sum_{i=1}^{n}(X_i - \overline{X})^2}\right)$
- $\hat{Y}_{(i)}$ l'estimateur de $Y_i$ ,calculé sans la ième observation.
- $s_{(i)}$ l'estimateur de $s$ l'écart-type résiduel, calculé sans la ième observation

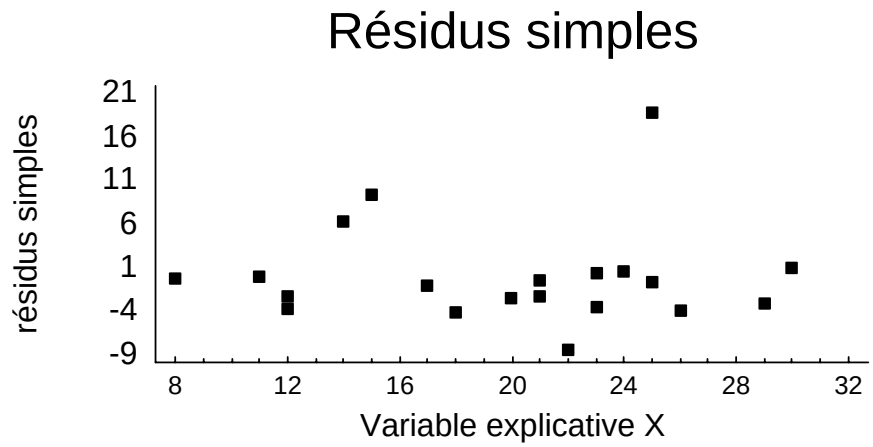
**Nous présentons ci-dessous différentes formules de résidus (rencontrés dans les logiciels ou la littérature) et leurs propriétés**.

***Les résidus simples***

$$e_i = Y_i - \hat{Y}_i \qquad \forall i \in \{1, n\}$$

Ces résidus sont les plus faciles à calculer et sont souvent donnés par les logiciels de régression. Ils s'expriment dans la même unité que la variable *Y.* Ils permettent éventuellement de repérer une valeur atypique mais ce ne sont pas les meilleurs pour cet objectif car ils n'ont pas la même variance et sont donc difficilement comparables.

⇒ Pour détecter les valeurs atypiques, il faut tracer le nuage de points des résidus simples croisés avec les valeurs de *X.* On peut ainsi repérer les formes suspectes de ce nuage qui devrait se répartir aléatoirement de part et d'autre de l'axe des abscisses.




Remarque : Pour un modèle avec constante, les résidus sont centrés sur zéro.

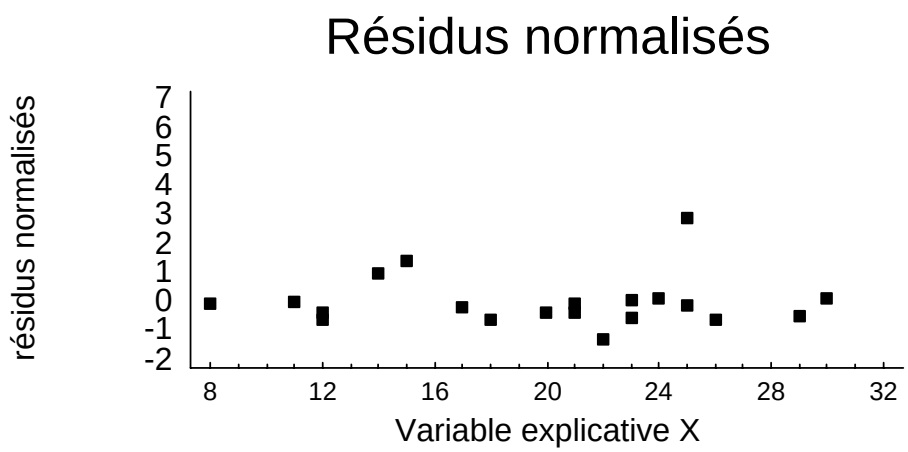
***Les résidus normalisés (ou partiels ou réduits)***

$$er_i = \frac{e_i}{s} \qquad \forall i \in \{1, n\} \text{ avec } s = \sqrt{\frac{\sum_{i=1}^{n} \left(Y_i - \hat{Y}_i\right)^2}{n-2}} = \sqrt{\frac{\sum_{i=1}^{n} e_i^2}{n-2}}$$

☐ Ces résidus sont sans unité. Dans la littérature, on identifie souvent la loi des résidus normalisés à une loi normale centrée réduite, ce qui n'est pas démontré théoriquement. De plus la variance des résidus normalisés n'est pas constante puisqu'elle vaut : $(1 - h_{ii})$ (voir tableau des résidus). Cependant plus *n* est grand, plus $h_{ii}$ est proche de zéro et donc plus la variance des résidus normalisés est proche de 1.
En général on admet que les résidus normalisés suivent une loi normale centrée réduite (si on a un nombre d'observations supérieur à 10), ce qui permet souvent de choisir un seuil critique simple de 2 à partir duquel on considère qu'un résidu est suspect. La valeur 2 correspond à un risque de 5% si l'hypothèse de normalité est vérifiée, c'est à dire qu'il y a environ une chance sur 20 qu'un résidu en valeur absolue se trouve au dessus de 2. On peut aussi prendre comme seuil la valeur 3 pour laquelle il y a une chance sur 100 qu'un résidu se trouve au-dessus en valeur absolue.

Remarque : l'utilitaire d'analyse d'Excel est censé calculer ces résidus mais en fait au lieu de diviser les résidus simples par l'estimateur de l'écart type résiduel ($s^2 = \frac{\sum_{i=1}^{n} e_i^2}{n-2}$), il les divise par l'écart type calculé d'après la série de résidus simples. (c'est à dire : $\frac{\sum_{i=1}^{n} e_i^2}{n-1}$)

⇒ Pour détecter les valeurs atypiques, il faut tracer le nuage de points des résidus normalisés croisés avec les valeurs de *X.* On peut ainsi repérer les formes suspectes de ce nuage qui devrait se répartir aléatoirement de part et d'autre de l'axe des abscisses. De plus les résidus doivent se trouver dans l'intervalle [-2 ;2], sinon ce sont des valeurs suspectes.

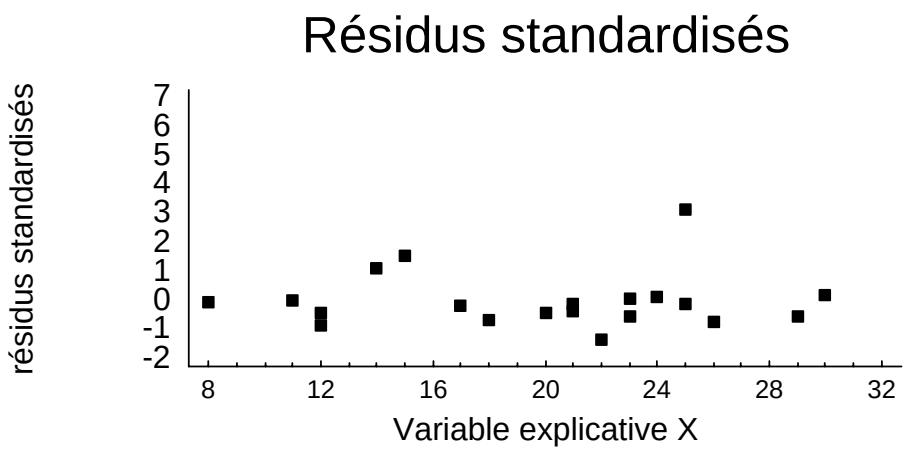



***Les résidus standardisés***

$$r_i = \frac{e_i}{s\sqrt{1-h_{ii}}} \qquad \forall i \in \{1,n\} \text{ avec} \qquad h_{ii} = \frac{1}{n} + \frac{(X_i - \overline{X})^2}{\sum_{i=1}^{n}(X_i - \overline{X})^2}$$

Les résidus standardisés ont les même propriétés que les résidus normalisés sauf que leur variance est constante et vaut exactement 1.

⇒ Pour détecter les valeurs atypiques, il faut tracer le nuage de points des résidus standardisés croisés avec les valeurs de *X.* On peut ainsi repérer les formes suspectes de ce nuage qui devrait se répartir aléatoirement de part et d'autre de l'axe des abscisses. De plus les résidus doivent se trouver dans l'intervalle [-2 ;2], sinon c'est que ce sont des valeurs suspectes.



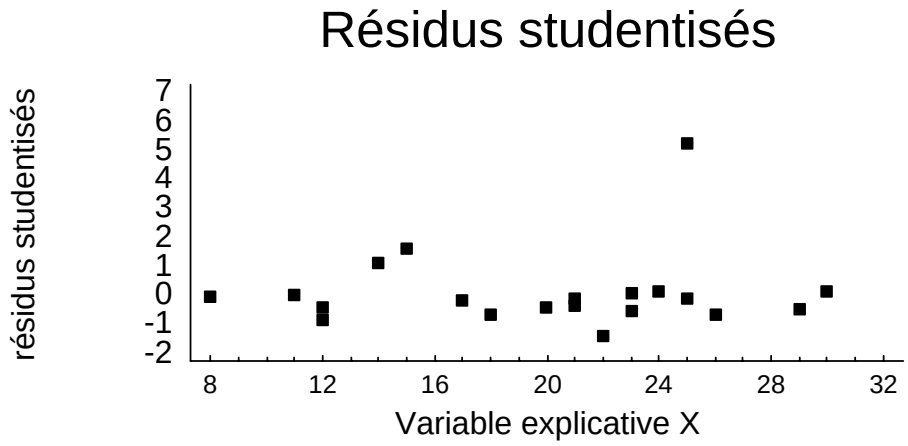

***Les résidus studentisés***

$$es_{(i)} = \frac{e_i}{s_{(i)}\sqrt{1-h_{ii}}} \qquad \forall i \in \{1, n\}$$

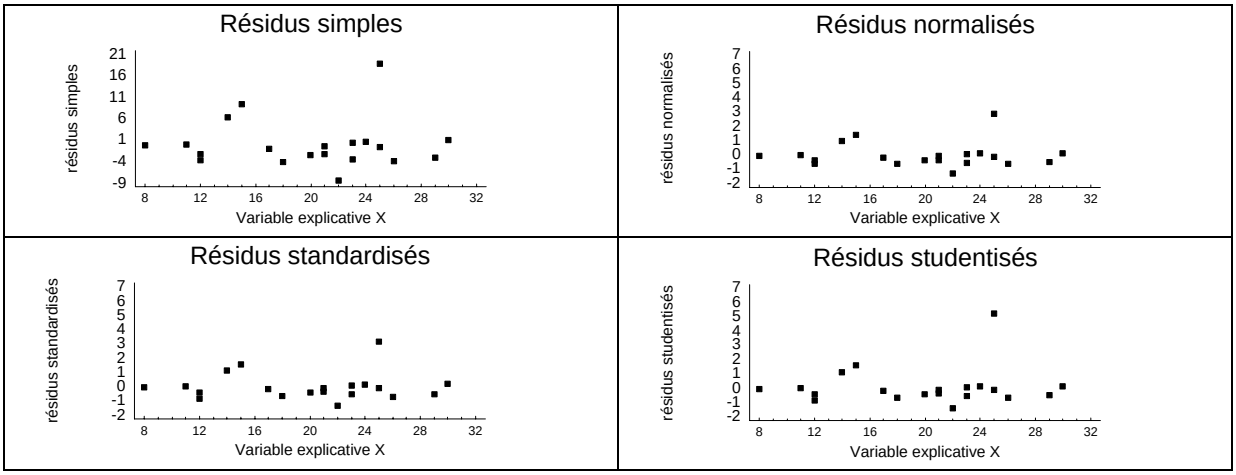
Les résidus studentisés permettent de mieux évaluer l'importance d'un résidu $e_i$ car on utilise une estimation de la variance résiduelle $\sigma$ sans utiliser l'observation i, notée $s_{(i)}$. Ainsi, enlever l'observation ayant un fort résidu $e_i$, conduit à une forte diminution de $\sigma$ et augmente donc la valeur du résidu studentisé. De plus, dans le cas où l'hypothèse de normalité des résidus est respectée, il est prouvé que les résidus studentisés suivent une loi de Student de degré de liberté *n-3* . On peut donc considérer qu'une observation *i* est mal reconstituée par le modèle dès que $\left|es_{(i)}\right| \geq t_{0.975}(n-3)$ .(valeur donnée dans la table de Student avec un risque de 5%). Cette valeur est proche de 2 quand le nombre d'observations *n* est suffisamment important, c'est pourquoi on prendra souvent 2 comme seuil critique pour les résidus studentisés.

$\Rightarrow$ Pour détecter les valeurs atypiques, il faut tracer le nuage de points des résidus studentisés croisés avec les valeurs de *X*. On peut ainsi repérer les formes suspectes de ce nuage qui devrait se répartir aléatoirement de part et d'autre de l'axe des abscisses. De plus les résidus doivent se trouver dans l'intervalle $\left[-t_{0975}(n-3);+t_{0.975}(n-3)\right]$ ( approximé à [-2 ;2]), sinon ce sont des valeurs suspectes.

Remarque : les résidus studentisés sont souvent appelés dans les logiciels ou la littérature « les Rstudent ».




Pour mieux comparer les résidus entre eux, réunissons côte à côte les graphes des différents résidus :




Dans cet exemple, les quatre types de résidus montrent qu'il existe une valeur atypique pour *X*=25, mais on peut remarquer que les résidus studentisés sont ceux qui la mettent le plus en évidence (la valeur du résidu studentisé est plus élevée).

## *Les valeurs influentes*

## *Les leviers*

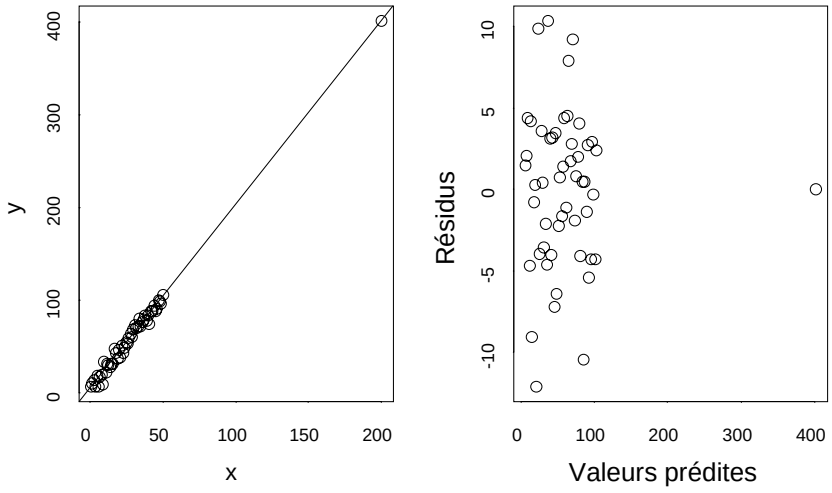
$$h_{ii} = \frac{1}{n} + \frac{(X_i - \overline{X})^2}{\sum_{i=1}^{n}(X_i - \overline{X})^2} \qquad \forall i \in \{1, n\}$$

Les élément diagonaux de la « hat matrice » (voir « annexe A ») notés $h_{ii}$ sont appelés les effets levier (en anglais « leverage »). Les observations pour lesquelles l'effet levier est important sont celles qui se situent loin de $\overline{X}$ et sont a priori peu nombreuses. Ces observations ont une grande influence sur le calcul de la droite de régression. Il n'est pas satisfaisant que les informations statistiques (estimations…) tiennent en grande partie à la présence d'un unique point (ou de peu de points) Le levier indique non seulement l'éloignement d'un point par rapport à $\overline{X}$, mais aussi l'impact ou l'importance que joue $Y_i$ dans l'estimation de $\hat{Y}_i$.

⇒ On peut considérer qu'un levier est important dès qu'il est supérieur à $\frac{4}{n}$.

⇒ Un effet levier important se détecte aussi en examinant les graphiques des données et des résidus.

# Effet levier important



## ***Critères d'influences***

Les critères qui suivent permettent de déceler des observations potentiellement influentes, ce qui n'est pas satisfaisant. Ils évaluent l'influence d'une observation sur certains paramètres (par exemple sur $\beta_0$, $\beta_1$ ou $Y$). Tous ces indicateurs proposent de comparer un paramètre estimé sans la ième observation et ce même paramètre estimé avec toutes les observations. Les plus utilisés sont présentés au début.

### ***La distance de Cook***

Elle mesure l'influence d'une observation sur l'ensemble des prévisions en prenant en compte l'effet levier et l'importance des résidus (d'où le dénominateur dans la formule). On la note $D_i$.

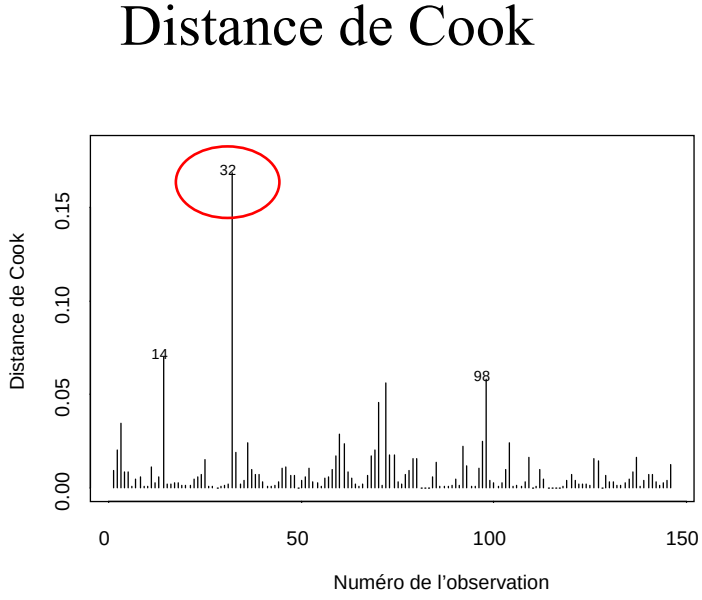
$$D_i = \frac{\sum_{j=1}^{n} (\hat{Y}_{(i)j} - \hat{Y}_j)^2}{2s^2} = \frac{h_{ii}}{2(1-h_{ii})} r_i^2 \qquad \text{pour} \quad i = 1,\ldots..n$$

avec :

- $\hat{Y}_{(i)j}$ prédiction de $Y_j$ calculée sans la ième observation $(X_i, Y_i)$
- $r_i$ résidu standardisé.

On considère qu'un point est atypique si la distance de Cook qui lui est associée est supérieure à $\frac{4}{n}$. (On trouve parfois dans les livres 1 comme seuil critique )

On peut aussi tracer le graphe des distances de Cook en fonction des numéros d'observations pour repérer rapidement la plus influente. La distance de Cook est le critère le plus souvent calculé par les logiciels de statistiques.



## Le DFITS

Il mesure l'influence de la ième observation $Y_i$ sur la prévision $\hat{Y}_i$ . Il est défini par :

$$DFITS_i = \frac{\hat{Y}_i - \hat{Y}_{(i)}}{s_{(i)}\sqrt{h_{ii}}}$$

avec :

- $Y_{(i)}$ l'estimateur de $Y$ pour la ième observation, mais sans utilisation de l'observation $i$
- $s_{(i)}$ l'estimateur de la variance résiduelle calculé sans la ième observation.
- $h_{ii}$ les leviers

On peut considérer que le DFITS est important dès qu'il est supérieur à $2\sqrt{\frac{2}{n}}$

## Le DFBETA

Il mesure l'influence de la ième observation sur l'estimation des coefficients $\beta_0$ et $\beta_1$. On le note $dfbetas_{i,j}$ :

$$dfbetas_{i,j} = \frac{b_j - b_{j(i)}}{s_{(i)}\sqrt{(X'X)^{-1}_{jj}}}$$

avec :

- $b_{j(i)}$ l'estimateur de $\beta_j$ obtenue sans l'observation i
- $s_{(i)}$ l'estimateur de la variance résiduelle calculé sans la ième observation
- $(X'X)^{-1}_{jj}$ le (j+1)ème élément diagonal de la matrice $(X'X)^{-1}$.

Le dfbeta est considéré comme suspect si sa valeur absolue dépasse $\frac{2}{\sqrt{n}}$

## Le COVRATIO

Le covratio mesure l'influence d'une observation sur la précision des estimateurs. Il est défini par :

$$\text{Covratio}_i = \frac{s^2_{(i)}\det(X'X)^{-1}_{(i)}}{s^2_{(i)}\det(X'X)^{-1}}$$

avec :

- $s_{(i)}$ l'estimateur de la variance résiduelle calculé sans la ième observation
- $\det(X'X)^{-1}_{(i)}$ le déterminant de la matrice $(X'X)^{-1}_{(i)}$ calculée sans la ième observation.

Un covratio plus grand que 1 indique que le fait de mettre l'observation augmente la qualité de l'estimation, alors qu'une plus petite valeur que 1 indique une diminution de la qualité de l'estimation.

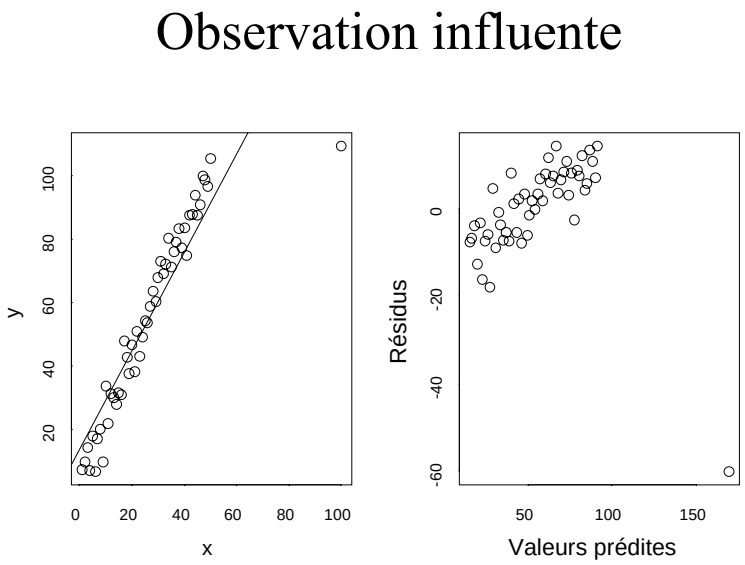
## ***Les résidus PRESS***

Les résidus PRESS permettent de repérer des points influents si on les compare au résidus simples. Ils mesurent en fait l'influence de la ième observation sur le ième résidu. Ils sont définis par :

$$e_{(i)} = Y_i - \hat{Y}_{(i)}$$

avec $\hat{Y}_{(i)}$ estimateur de $Y_i$ calculé sans la ième observation.

*Graphiquement, on repère aussi les observations influentes en traçant le graphe des résidus en fonction des valeurs prédites.*



## ***RESUME DES CRITERES***

| *Critères de détection de valeurs atypiques* | *Seuil critique à partir duquel l'observation est considérée suspecte* |
|---|---|
| **\|Résidu studentisé \| ou \|Rstudent\|** | >2 |
| **\|Levier\|** | $> \frac{4}{n}$ |
| **\|Distance de Cook\|** | $> \frac{4}{n}$ |
| **\|DFITS\|** | $> 2\sqrt{\frac{2}{n}}$ |
| **\|DFBETAS\|** | $> \frac{2}{\sqrt{n}}$ |
| **\|COVRATIO\|** | >1 augmente la qualité de l'estimation<br><1 diminue la qualité de l'estimation |

## TABLEAU RECAPITULATIF SUR LES RESIDUS

| **RESIDUS**<br>*RAW RESIDUALS* | **RESIDUS PARTIELS ou REDUITS ou NORMALISES**<br>*SEMI-STUDENTIZED RESIDUALS* | **RESIDUS STANDARDISES**<br>*INTERNALLY STUDENTIZED RESIDUALS* | **RESIDUS PRESS**<br>PRESS RESIDUALS | **RESIDUS STUDENTISES**<br>*EXTERNALLY STUDENTIZED RESIDUALS ou Rstudent* |
|---|---|---|---|---|
| $e_i = Y_i - \hat{Y}_i \quad \forall i \in \{1,n\}$ | $er_i = \frac{e_i}{s} \quad \forall i \in \{1,n\}$ | $r_i = \frac{e_i}{s\sqrt{1-h_{ii}}} \quad \forall i \in \{1,n\}$ | $e_{(i)} = Y_i - \hat{Y}_{(i)} \quad \forall i \in \{1,n\}$ | $es_{(i)} = \frac{e_i}{s_{(i)}\sqrt{1-h_{ii}}} \quad \forall i \in \{1,n\}$ |
| Les résidus ne sont pas indépendants | | | | |
| Les résidus ne sont pas identiquement distribués Ils suivent tous une loi normale mais n'ont pas même variance. | Les résidus ne sont pas identiquement distribués. Il n'est pas démontré qu'ils suivent une loi normale. Ils n'ont pas même variance. | Les résidus ne sont pas identiquement distribués. Il n'est pas démontré qu'ils suivent une loi normale. Ils ont tous même variance. | Les résidus ne sont pas identiquement distribués. Il n'est pas démontré qu'ils suivent une loi normale. Ils n'ont pas même variance. | Les résidus sont identiquement distribués. Ils suivent tous une loi de Student *t*(n-3) et ont même variance. |
| $Var\, e_i = \sigma^2(1-h_{ii})$ $\forall i \in \{1,n\}$ | $Var\, er_i \approx (1-h_{ii})$ $\forall i \in \{1,n\}$ | $Var(r_i) = 1 \ \forall i \in \{1,n\}$ | $Var\, e_{(i)} = \frac{\sigma^2}{1-h_{ii}} \quad \forall i \in \{1,n\}$ | $Var\, es_{(i)} = \frac{n-3}{n-5}$ |
| <u>*Remarque :*</u> si l'hypothèse de normalité est vérifiée, ils suivent une loi normale $N(0,\sigma\sqrt{1-h_{ii}})$ | <u>*Remarque*</u> : si l'hypothèse de normalité est vérifiée, on considère souvent qu'ils suivent une loi normale centrée réduite N(0,1) et qu'ils sont identiquement distribués. | <u>*Remarque*</u> : on peut mieux comparer les résidus car ils ont tous même variance. On considère qu'ils suivent une loi N(0,1) | <u>*Remarque*</u>: ils permettent de repérer les points influents si on calcule la différence $e_i - e_{(i)}$. | <u>*Remarque*</u> : ils permettent de trouver les points atypiques (outliers). Ils suivent une loi de Student de degré de liberté n-3 |

# REGRESSION LINEAIRE SIMPLE

## HOMOSCEDASTICITE ou égalité des variances

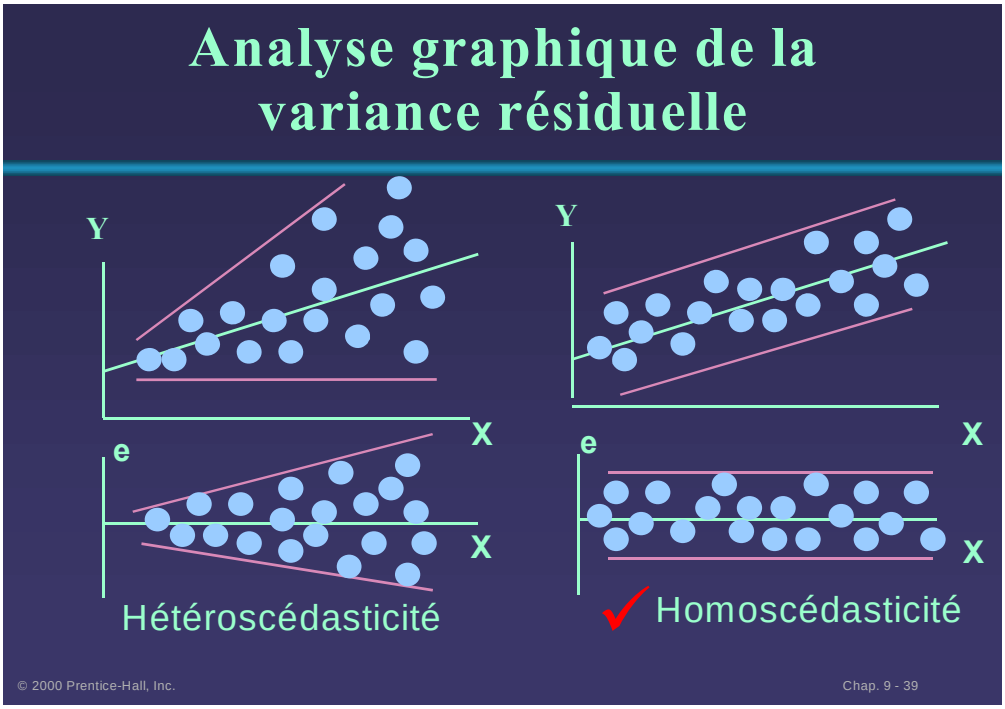
Nous avons vu dans la fiche «Hypothèses» qu'il faut vérifier l'égalité des variances des erreurs (homoscédasticité). Cela signifie que toutes les distributions de $Y$ (les $Y_i$) doivent avoir le même écart-type : la variance résiduelle est constante sur le domaine étudié.

### *Les graphiques*

Un graphique permet de voir dans un premier temps si il semble y avoir homoscédasticité ou pas. On peut visualiser le graphique de $Y$ la variable expliquée en fonction de $X$ la variable explicative ou mieux utiliser le graphique des résidus. (voir fiche RESIDUS)

Remarque :

L'hétéroscedasticité est un des cas les plus fréquents de non respect des hypothèses. Si la variance n'est pas constante sur tout le domaine, utiliser les "moindres carrés" avec éventuellement l'inverse de la variance pour pondération. (voir fiche "REGRESSION PONDEREE")



## ***Les tests.***

Il est préférable pour utiliser les tests d'introduire les notations suivantes :

***Les données :***

Pour obtenir sa série de données, l'expérimentateur procède de la façon suivante :

il observe une ou plusieurs valeurs que prend la variable *Y* pour différentes valeurs fixées de *X* appelées « niveaux de la variable *X* » :

$\Rightarrow$ On note les niveaux de *X* : $x_1, x_2, \ldots\ldots\ldots x_p$

$\downarrow$

A chaque niveau $x_i$ ($i \in \{1, p\}$) de *X*, il note les $n_i$ valeurs de *Y* qu'il observe (si il décide de répéter $n_i$ fois l'expérience).

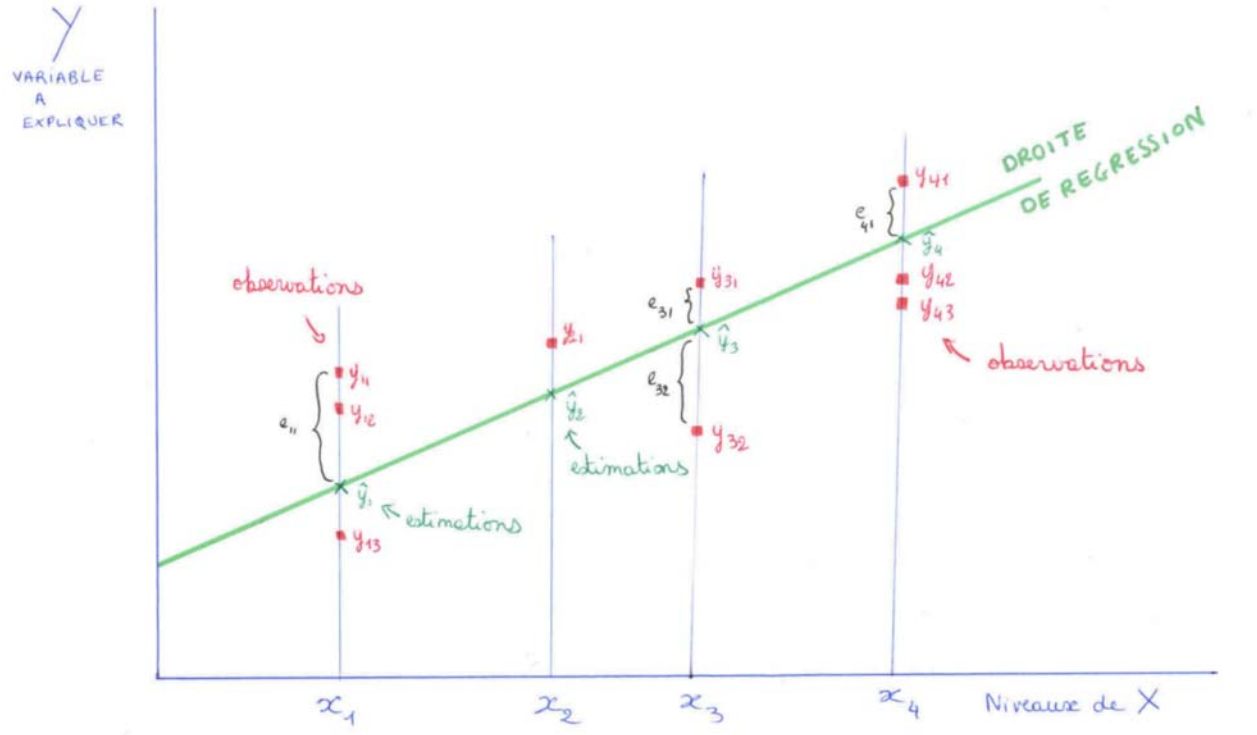
$\Rightarrow$ On note ces $n_i$ valeurs prises par *Y* quand *X* vaut $x_i$ : $y_{i1}, y_{i2}, \ldots\ldots\ldots\ldots, y_{in_i}$

$\downarrow$

$\Rightarrow$ Les résidus associées au $n_i$ couples d'observations $(x_i, y_{i1}), (x_i, y_{i2}), \ldots\ldots\ldots\ldots (x_i, y_{in_i})$ du niveau $x_i$ sont notées : $e_{i1}, e_{i2}, \ldots\ldots\ldots e_{in_i}$

Exemples pour comprendre les notations :

1)



| niveaux de X | *X* :concentration en nitrate du produit | Valeurs de *Y* | *Y* : réponse de l'analyseur | Remarques |
|---|---|---|---|---|
| $x_1$ | 199 | $y_{11}$ | 945 | $y_{11}, y_{12}, y_{13}, y_{14}$ sont des réalisations de $Y_1$. L'expérimentateur a fait 4 répétitions : $n_1 = 4$ |
| $x_1$ | 199 | $y_{12}$ | 965 | |
| $x_1$ | 199 | $y_{13}$ | 978 | |
| $x_1$ | 199 | $y_{14}$ | 960 | |
| $x_2$ | 210,5 | $y_{21}$ | 1059 | $y_{21}, y_{22}, y_{23}, y_{24}$ sont des réalisations de $Y_2$. L'expérimentateur a fait 4 répétitions : $n_2 = 4$ |
| $x_2$ | 210,5 | $y_{22}$ | 1025 | |
| $x_2$ | 210,5 | $y_{23}$ | 1015 | |
| $x_2$ | 210,5 | $y_{24}$ | 1042 | |
| $x_3$ | 224 | $y_{31}$ | 1056 | $y_{31}, y_{32}$ sont des réalisations de $Y_3$. L'expérimentateur a fait 2 répétitions : $n_3 = 2$ |
| $x_3$ | 224 | $y_{32}$ | 1075 | |
| $x_4$ | 233 | $y_{41}$ | 1165 | $y_{41}, y_{42}$ sont des réalisations de $Y_4$. L'expérimentateur a fait 2 répétitions : $n_4 = 2$ |
| $x_4$ | 233 | $y_{42}$ | 1179 | |
| $x_5$ | 260 | $y_{51}$ | 1281 | $y_{51}, y_{52}, y_{53}, y_{54}$ sont des réalisations de $Y_5$. L'expérimentateur a fait 4 répétitions : $n_5 = 4$ |
| $x_5$ | 260 | $y_{52}$ | 1292 | |
| $x_5$ | 260 | $y_{53}$ | 1275 | |
| $x_5$ | 260 | $y_{54}$ | 1299 | |

## ***Les tests***

*ATTENTION : les tests ne sont valables que sous l'hypothèse de la normalité des erreurs.*

Rappel : on veut vérifier à l'aide des données que pour chaque niveau de *X*, les distributions de *Y* ont même écart-type.

---

On appelle $s_1^2, s_2^2 ..., s_p^2$ les variances estimées de *Y* aux niveaux $\{x_1, x_2, .....x_p\}$ de *X*.

Rappelons que $s_i^2 = \frac{1}{n_i - 1} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2$ avec $\bar{y}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} y_{ij} \quad \forall i \in \{1, p\}$

---

Plusieurs tests statistiques permettent de tester l'homoscédasticité :

## Le test de Fisher :

Le test de Fisher s'applique pour tester l'égalité de deux variances uniquement ($p$ = 2).

⇒ Il est décrit dans « Méthodes statistiques pour l'ingénieur » de Olivier Gaudoin p66. Ce polycopié est disponible dans le dossier DOCUMENTS INTERNET sous le nom de « Gaudoin » ou à l'adresse suivante :

http://www-lmc.imag.fr/lmc-sms/Olivier.Gaudoin/modlin.pdf

## Le test de Cochran : (homogénéité des variances)

Le test de Cochran permet de contrôler l'égalité de $p$ variances en suspectant la variance la plus grande, à condition que la population soit normale et que toutes les variances soient estimées par un même nombre de valeurs $n$ (le nombre de degrés de liberté associé à chacune des estimations de ces variances doit être constant).

Soit $p$ variances : $s_1^2, s_2^2 ..., s_p^2$

On note : $s_{\max}^2 = \max\left\{s_1^2, s_2^2 ..., s_p^2\right\}$

La statistique $C$ du test de Cochran est

$$C = \frac{s_{\max}^2}{\sum_{i=1}^{p} s_i^2}$$

où $s_{\max}$ est l'écart-type le plus élevé de l'ensemble.

La table du test de Cochran fournit une valeur $C(n, p, \alpha)$ en fonction du nombre $n$ de données, du nombre de variances à comparer $p$, et de la valeur du risque $\alpha$.

Après avoir fixé la valeur du risque $\alpha$ , on conclut au rejet ou non-rejet de l'homogénéité des variances de la façon suivante :

1. Si la statistique C du test (calculée avec les données) est inférieure ou égale à $C(n, p, \alpha)$ l'hypothèse d'homogénéité des variances est acceptée
2. Si la statistique C du test est supérieure à $C(n, p, \alpha)$, l'hypothèse d'homogénéité des variances est rejetée.

**Note :** $\alpha$ est souvent fixé à 5 % ou à 1 %

| |
|---|
| **$\alpha$=0.05** |

| p \ n | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|
| 2 | 0.999 | 0.975 | 0.939 | 0.906 | 0.877 | 0,853 | 0.833 | 0.816 | 0.801 |
| 3 | 0.967 | 0.871 | 0.798 | 0.746 | 0.707 | 0,677 | 0.653 | 0.633 | 0.617 |
| 4 | 0.907 | 0.768 | 0.684 | 0.629 | 0.590 | 0,560 | 0.537 | 0.518 | 0.502 |
| 5 | 0.841 | 0.684 | 0.598 | 0.544 | 0.507 | 0,478 | 0.456 | 0.438 | 0.424 |
| 6 | 0.781 | 0.616 | 0.532 | 0.480 | 0.445 | 0,418 | 0.398 | 0.382 | 0.368 |
| 7 | 0.727 | 0.561 | 0.480 | 0.431 | 0.397 | 0,373 | 0.354 | 0.338 | 0.326 |
| 8 | 0.680 | 0.516 | 0.438 | 0.391 | 0.360 | 0,336 | 0.319 | 0.304 | 0.293 |
| 9 | 0.639 | 0.478 | 0.403 | 0.358 | 0.329 | 0,307 | 0.290 | 0.277 | 0.266 |
| 10 | 0.602 | 0.445 | 0.373 | 0.331 | 0.303 | 0,282 | 0.267 | 0.254 | 0.244 |
| 12 | 0.541 | 0.392 | 0.326 | 0.288 | 0.262 | 0,244 | 0.230 | 0.219 | 0.210 |
| 15 | 0.471 | 0.335 | 0.276 | 0.242 | 0.220 | 0,203 | 0.191 | 0.182 | 0.174 |
| 20 | 0.389 | 0.271 | 0.221 | 0.192 | 0.174 | 0,160 | 0.150 | 0.142 | 0.136 |
| 24 | 0.343 | 0.235 | 0.191 | 0.166 | 0.149 | 0,137 | 0.129 | 0.122 | 0.116 |
| 30 | 0.293 | 0.198 | 0.159 | 0.138 | 0.124 | 0,114 | 0.106 | 0.100 | 0.096 |
| 40 | 0.237 | 0.158 | 0.126 | 0.108 | 0.097 | 0,089 | 0.083 | 0.078 | 0.075 |

$\alpha$ = 0.01

| p \ n | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|
| 2 | — | 0.995 | 0.979 | 0.959 | 0.937 | 0,917 | 0.899 | 0.882 | 0.867 |
| 3 | 0.993 | 0.942 | 0.883 | 0.834 | 0.793 | 0,761 | 0.734 | 0.711 | 0.691 |
| 4 | 0.968 | 0.864 | 0.781 | 0.721 | 0.676 | 0,641 | 0.613 | 0.590 | 0.570 |
| 5 | 0.928 | 0.789 | 0.696 | 0.633 | 0.588 | 0,553 | 0.526 | 0.504 | 0.485 |
| 6 | 0.883 | 0.722 | 0.626 | 0.564 | 0.520 | 0,487 | 0.461 | 0.440 | 0.423 |
| 7 | 0.838 | 0.664 | 0.569 | 0.508 | 0.466 | 0,435 | 0.411 | 0.391 | 0.375 |
| 8 | 0.795 | 0.615 | 0.521 | 0.463 | 0.423 | 0,393 | 0.370 | 0.352 | 0.337 |
| 9 | 0.754 | 0.573 | 0.481 | 0.425 | 0.387 | 0,359 | 0.338 | 0.321 | 0.307 |
| 10 | 0.718 | 0.536 | 0.447 | 0.393 | 0.357 | 0,331 | 0.311 | 0.295 | 0.281 |
| 12 | 0.653 | 0.475 | 0.392 | 0.343 | 0.310 | 0,286 | 0.268 | 0.254 | 0.242 |
| 15 | 0.575 | 0.407 | 0.332 | 0.288 | 0.259 | 0,239 | 0.223 | 0.210 | 0.200 |
| 20 | 0.480 | 0.330 | 0.265 | 0.229 | 0.205 | 0,188 | 0.175 | 0.165 | 0.157 |
| 24 | 0.425 | 0.287 | 0.230 | 0.197 | 0.176 | 0,161 | 0.150 | 0.141 | 0.134 |
| 30 | 0.363 | 0.241 | 0.191 | 0.164 | 0.145 | 0,133 | 0.123 | 0.116 | 0.110 |
| 40 | 0.294 | 0.192 | 0.151 | 0.128 | 0.114 | 0,103 | 0.096 | 0.090 | 0.085 |

**α =0.01**

***Table des valeurs C(n,p,$\alpha$) pour le test de Cochran***

## Le test de Hartley :

Le test de Hartley est utilisé pour contrôler l'égalité de *p* variances dans le cas de populations normales. Toutes les variances doivent être estimées par un même nombre de valeurs (le nombre de degrés de liberté associé à chacune des estimations de ces variances est constant).

Soit *p* variances : $s_1^2, s_2^2 ..., s_p^2$

On note : $s_{\max}^2 = \max\left\{s_1^2, s_2^2 ..., s_p^2\right\}$ et $s_{\min}^2 = \min\left\{s_1^2, s_2^2 ..., s_p^2\right\}$

La statistique $C$ du test de Hartley est :

$$C = \frac{s_{\max}^2}{s_{\min}^2}$$

où $s_{\max}^2$ est la plus forte des estimations de variances, $s_{\min}^2$ la plus faible.

La table du test de Hartley fournit une valeur C($\nu$, p, $\alpha$) en fonction du nombre de degrés de liberté $\nu$ (nombre de données – 1), du nombre de variances à comparer p, et de la valeur du risque $\alpha$.

Après avoir fixé la valeur du risque $\alpha$ , on conclut au rejet ou non-rejet de l'homogénéité des variances de la façon suivante :

3. Si la statistique C du test (calculée avec les données) est inférieure ou égale à *C(n, p, $\alpha$)* l'hypothèse d'homogénéité des variances est acceptée
4. Si la statistique C du test est supérieure à *C(n, p, $\alpha$)*, l'hypothèse d'homogénéité des variances est rejetée.

## TABLE DE HARTLEY

| ν \ p | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | k \ ν |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 2 | 39,0 | 87,5 | 142 | 202 | 266 | 333 | 403 | 475 | 550 | 626 | 704 | 2 |
| 3 | 15,4 | 27,8 | 39,2 | 50,7 | 62,0 | 72,9 | 83,5 | 93,9 | 104 | 114 | 124 | 3 |
| 4 | 9,60 | 15,5 | 20,6 | 25,2 | 29,5 | 33,6 | 37,5 | 41,1 | 44,6 | 48,0 | 51,4 | 4 |
| 5 | 7,15 | 10,8 | 13,7 | 16,3 | 18,7 | 20,8 | 22,9 | 24,7 | 26,5 | 28,2 | 29,9 | 5 |
| 6 | 5,82 | 8,38 | 10,4 | 12,1 | 13,7 | 15,0 | 16,3 | 17,5 | 18,6 | 19,7 | 20,7 | 6 |
| 7 | 4,99 | 6,94 | 8,44 | 9,70 | 10,8 | 11,8 | 12,7 | 13,5 | 14,3 | 15,1 | 15,8 | 7 |
| 8 | 4,43 | 6,00 | 7,18 | 8,12 | 9,03 | 9,78 | 10,5 | 11,1 | 11,7 | 12,2 | 12,7 | 8 |
| 9 | 4,03 | 5,34 | 6,31 | 7,11 | 7,80 | 8,41 | 8,95 | 9,45 | 9,91 | 10,3 | 10,7 | 9 |
| 10 | 3,72 | 4,85 | 5,67 | 6,34 | 6,92 | 7,42 | 7,87 | 8,28 | 8,66 | 9,01 | 9,34 | 10 |
| 11 | 3,28 | 4,16 | 4,79 | 5,30 | 5,72 | 6,09 | 6,42 | 6,72 | 7,00 | 7,25 | 7,48 | 11 |
| 15 | 2,86 | 3,54 | 4,01 | 4,37 | 4,68 | 4,95 | 5,19 | 5,40 | 5,59 | 5,77 | 5,93 | 15 |
| 20 | 2,46 | 2,95 | 3,29 | 3,54 | 3,76 | 3,94 | 4,10 | 4,24 | 4,37 | 4,49 | 4,59 | 20 |
| 30 | 2,07 | 2,40 | 2,61 | 2,78 | 2,91 | 3,02 | 3,12 | 3,21 | 3,29 | 3,36 | 3,39 | 30 |
| 60 | 1,67 | 1,85 | 1,96 | 2,04 | 2,11 | 2,17 | 2,22 | 2,26 | 2,30 | 2,33 | 2,36 | 60 |
| ∞ | 1,00 | 1,00 | 1,00 | 1,00 | 1,00 | 1,00 | 1,00 | 1,00 | 1,00 | 1,00 | 1,00 | ∞ |
| ν \ k | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | k \ ν |

$\alpha = 0,01$

| ν \ p | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | k \ ν |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 2 | 199 | 448 | 729 | 1 036 | 1 362 | 1 705 | 2 063 | 2 432 | 2 813 | 3 204 | 3 605 | 2 |
| 3 | 47,5 | 85 | 120 | 151 | 184 | 216 | 249 | 281 | 310 | 337 | 361 | 3 |
| 4 | 23,2 | 37 | 49 | 59 | 69 | 79 | 89 | 97 | 106 | 113 | 120 | 4 |
| 5 | 14,9 | 22 | 28 | 33 | 38 | 42 | 46 | 50 | 54 | 57 | 60 | 5 |
| 6 | 11,1 | 15,5 | 19,1 | 22 | 25 | 27 | 30 | 32 | 34 | 36 | 37 | 6 |
| 7 | 8,89 | 12,1 | 14,5 | 16,5 | 18,4 | 20 | 22 | 23 | 24 | 26 | 27 | 7 |
| 8 | 7,50 | 9,9 | 11,7 | 13,2 | 14,5 | 15,8 | 16,9 | 17,9 | 18,9 | 19,8 | 21 | 8 |
| 9 | 6,54 | 8,5 | 9,9 | 11,1 | 12,1 | 13,1 | 13,9 | 14,7 | 15,3 | 16,0 | 16,6 | 9 |
| 10 | 5,85 | 7,4 | 8,6 | 9,6 | 10,4 | 11,1 | 11,8 | 12,4 | 12,9 | 13,4 | 13,9 | 10 |
| 11 | 4,91 | 6,1 | 6,9 | 7,6 | 8,2 | 8,7 | 9,1 | 9,5 | 9,9 | 10,2 | 10,6 | 11 |
| 15 | 4,07 | 4,9 | 5,5 | 6,0 | 6,4 | 6,7 | 7,1 | 7,3 | 7,5 | 7,8 | 8,0 | 15 |
| 20 | 3,32 | 3,8 | 4,3 | 4,6 | 4,9 | 5,1 | 5,3 | 5,5 | 5,6 | 5,8 | 5,9 | 20 |
| 30 | 2,63 | 3,0 | 3,3 | 3,4 | 3,6 | 3,7 | 3,8 | 3,9 | 4,0 | 4,1 | 4,2 | 30 |
| 60 | 1,96 | 2,2 | 2,3 | 2,4 | 2,4 | 2,5 | 2,5 | 2,6 | 2,6 | 2,7 | 2,7 | 60 |
| ∞ | 1,00 | 1,0 | 1,0 | 1,0 | 1,0 | 1,0 | 1,0 | 1,0 | 1,0 | 1,0 | 1,0 | ∞ |
| ν \ k | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | k \ ν |

$\alpha$=0.05

Remarque : il existe deux autres tests qui permettent de tester l'homogénéité des variances :

⇒ ***le test de Levene***
Il est décrit dans le document Internet « Régression 1 » note 3 page 7 disponible dans le dossier DOCUMENTS INTERNET ou à l'adresse suivante :
http://www.stat.psu.edu/~cspark/stat501/

⇒ ***le test de Bartlett*** expliqué dans « théorie et méthodes statistiques » de Pierre Dagnelie p54 (disponible au service statistiques)

# REGRESSION LINEAIRE SIMPLE

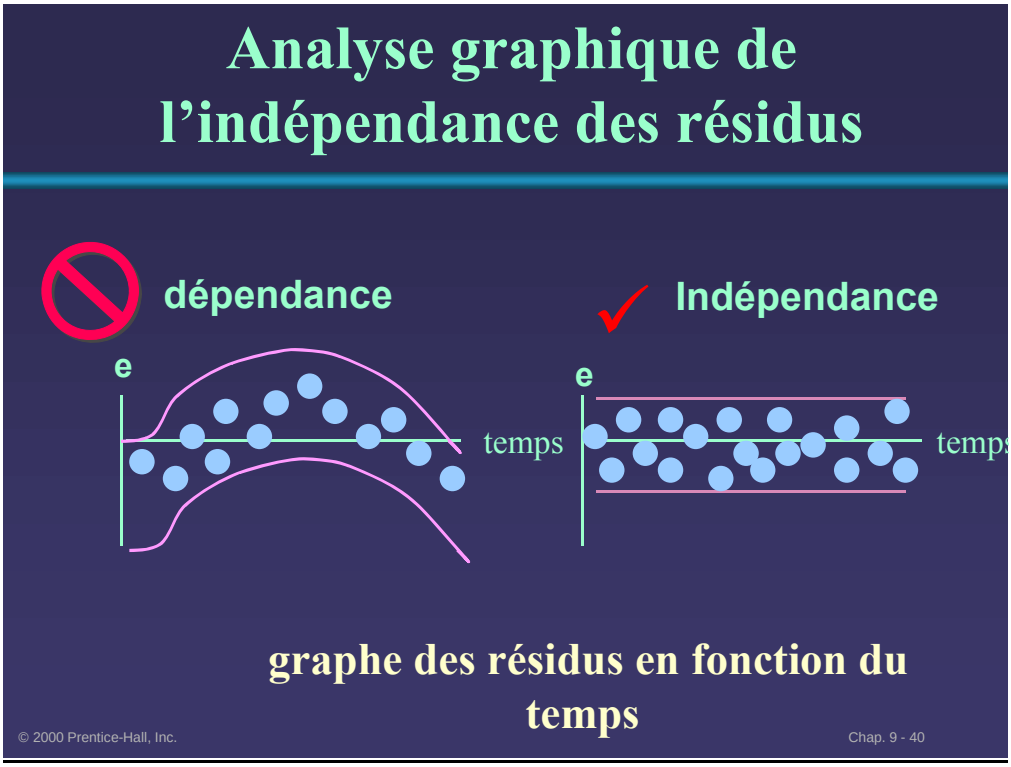
## Corrélation des termes d'erreurs

On veut vérifier qu'il n'y a aucun lien entre l'erreur commise sur une mesure et les mesures effectuées précédemment. Pour cela, deux méthodes sont possibles : l'étude graphique ou les tests.

### 1. *Etude graphique*

On trace le graphe des résidus en fonction du temps (voir fiche « résidus ») et on regarde si ils se répartissent de manière aléatoire autour de l'axe des abscisses. Si ce n'est pas le cas, on peut mettre en doute la non-corrélation des termes d'erreur.

Exemple :




### 2. *Les tests*

Pour détecter un problème de corrélation des erreurs dans une régression, on peut utiliser le test de Durbin-Watson. Il s'utilise dans le cas où *Y* suit une loi normale (voir fiche "NORMALITE".).
On utilise les résidus $\{e_1, e_2, ....e_n\}$ pour calculer le coefficient de Durbin-Watson qui permet de prendre une décision quant à la non-corrélation des erreurs.

Le coefficient de Durbin-Watson est défini ainsi :

$$d = \frac{\sum_{i=1}^{n-1}(e_{i+1} - e_i)^2}{\sum_{1}^{n} e_i^2}$$

La table de Durbin-Watson donne deux valeurs $d_1$ et $d_2$ en fonction du nombre d'observations n, du nombre de variables explicatives *p* (*p*=1 dans le cas de la régression simple) et du risque $\alpha$.
Après avoir fixé la valeur du risque $\alpha$, on conclut au rejet ou non-rejet de non corrélation des erreurs de la façon suivante :

5. Si le coefficient de Durbin-Watson d (calculé avec les données) est inférieure à $d_1$, l'hypothèse de non corrélation des erreurs est rejetée.
6. Si le coefficient de Durbin-Watson d (calculé avec les données) est supérieur à $d_2$, l'hypothèse de non corrélation des erreurs est acceptée .
7. Si le coefficient de Durbin-Watson d (calculé avec les données) est compris entre $d_1$ et $d_2$, il y a un doute.(on ne peut pas conclure)

**TABLE DE DURBIN-WATSON**

$\alpha = 0{,}05$

| n \ p | 1 | | 2 | | 3 | | 4 | | 5 | |
|---|---|---|---|---|---|---|---|---|---|---|
| | $d_1$ | $d_2$ | $d_1$ | $d_2$ | $d_1$ | $d_2$ | $d_1$ | $d_2$ | $d_1$ | $d_2$ |
| 15 | 1,08 | 1,36 | 0,95 | 1,54 | 0,82 | 1,75 | 0,69 | 1,97 | 0,56 | 2,21 |
| 16 | 1,10 | 1,37 | 0,98 | 1,54 | 0,86 | 1,73 | 0,74 | 1,93 | 0,62 | 2,15 |
| 17 | 1,13 | 1,38 | 1,02 | 1,54 | 0,90 | 1,71 | 0,78 | 1,90 | 0,67 | 2,10 |
| 18 | 1,16 | 1,39 | 1,05 | 1,53 | 0,93 | 1,69 | 0,82 | 1,87 | 0,71 | 2,06 |
| 19 | 1,18 | 1,40 | 1,08 | 1,53 | 0,97 | 1,68 | 0,86 | 1,85 | 0,75 | 2,02 |
| 20 | 1,20 | 1,41 | 1,10 | 1,54 | 1,00 | 1,68 | 0,90 | 1,83 | 0,79 | 1,99 |
| 21 | 1,22 | 1,42 | 1,13 | 1,54 | 1,03 | 1,67 | 0,93 | 1,81 | 0,83 | 1,96 |
| 22 | 1,24 | 1,43 | 1,15 | 1,54 | 1,05 | 1,66 | 0,96 | 1,80 | 0,86 | 1,94 |
| 23 | 1,26 | 1,44 | 1,17 | 1,54 | 1,08 | 1,66 | 0,99 | 1,79 | 0,90 | 1,92 |
| 24 | 1,27 | 1,45 | 1,19 | 1,55 | 1,10 | 1,66 | 1,01 | 1,78 | 0,93 | 1,90 |
| 25 | 1,29 | 1,45 | 1,21 | 1,55 | 1,12 | 1,66 | 1,04 | 1,77 | 0,95 | 1,89 |
| 26 | 1,30 | 1,46 | 1,22 | 1,55 | 1,14 | 1,65 | 1,06 | 1,76 | 0,98 | 1,88 |
| 27 | 1,32 | 1,47 | 1,24 | 1,56 | 1,16 | 1,65 | 1,08 | 1,76 | 1,01 | 1,86 |
| 28 | 1,33 | 1,48 | 1,26 | 1,56 | 1,18 | 1,65 | 1,10 | 1,75 | 1,03 | 1,85 |
| 29 | 1,34 | 1,48 | 1,27 | 1,56 | 1,20 | 1,65 | 1,12 | 1,74 | 1,05 | 1,84 |
| 30 | 1,35 | 1,49 | 1,28 | 1,57 | 1,21 | 1,65 | 1,14 | 1,74 | 1,07 | 1,83 |
| 35 | 1,40 | 1,52 | 1,34 | 1,58 | 1,28 | 1,65 | 1,22 | 1,73 | 1,16 | 1,80 |
| 40 | 1,44 | 1,54 | 1,39 | 1,60 | 1,34 | 1,66 | 1,29 | 1,72 | 1,23 | 1,79 |
| 45 | 1,48 | 1,57 | 1,43 | 1,62 | 1,38 | 1,67 | 1,34 | 1,72 | 1,29 | 1,78 |
| 50 | 1,50 | 1,59 | 1,46 | 1,63 | 1,42 | 1,67 | 1,38 | 1,72 | 1,34 | 1,77 |
| 60 | 1,55 | 1,62 | 1,51 | 1,65 | 1,48 | 1,69 | 1,44 | 1,73 | 1,41 | 1,77 |
| 70 | 1,58 | 1,64 | 1,55 | 1,67 | 1,52 | 1,70 | 1,49 | 1,74 | 1,46 | 1,77 |
| 80 | 1,61 | 1,66 | 1,59 | 1,69 | 1,56 | 1,72 | 1,53 | 1,74 | 1,51 | 1,77 |
| 90 | 1,63 | 1,68 | 1,61 | 1,70 | 1,59 | 1,73 | 1,57 | 1,75 | 1,54 | 1,78 |
| 100 | 1,65 | 1,69 | 1,63 | 1,72 | 1,61 | 1,74 | 1,59 | 1,76 | 1,57 | 1,78 |

$\alpha = 0{,}01$

| n \ p | 1 | | 2 | | 3 | | 4 | | 5 | |
|---|---|---|---|---|---|---|---|---|---|---|
| | $d_1$ | $d_2$ | $d_1$ | $d_2$ | $d_1$ | $d_2$ | $d_1$ | $d_2$ | $d_1$ | $d_2$ |
| 15 | 0,81 | 1,07 | 0,70 | 1,25 | 0,59 | 1,46 | 0,49 | 1,70 | 0,39 | 1,96 |
| 16 | 0,84 | 1,09 | 0,74 | 1,25 | 0,63 | 1,44 | 0,53 | 1,66 | 0,44 | 1,90 |
| 17 | 0,87 | 1,10 | 0,77 | 1,25 | 0,67 | 1,43 | 0,57 | 1,63 | 0,48 | 1,85 |
| 18 | 0,90 | 1,12 | 0,80 | 1,26 | 0,71 | 1,42 | 0,61 | 1,60 | 0,52 | 1,80 |
| 19 | 0,93 | 1,13 | 0,83 | 1,26 | 0,74 | 1,41 | 0,65 | 1,58 | 0,56 | 1,77 |
| 20 | 0,95 | 1,15 | 0,86 | 1,27 | 0,77 | 1,41 | 0,68 | 1,57 | 0,60 | 1,74 |
| 21 | 0,97 | 1,16 | 0,89 | 1,27 | 0,80 | 1,41 | 0,72 | 1,55 | 0,63 | 1,71 |
| 22 | 1,00 | 1,17 | 0,91 | 1,28 | 0,83 | 1,40 | 0,75 | 1,54 | 0,66 | 1,69 |
| 23 | 1,02 | 1,19 | 0,94 | 1,29 | 0,86 | 1,40 | 0,77 | 1,53 | 0,70 | 1,67 |
| 24 | 1,04 | 1,20 | 0,96 | 1,30 | 0,88 | 1,41 | 0,80 | 1,53 | 0,72 | 1,66 |
| 25 | 1,05 | 1,21 | 0,98 | 1,30 | 0,90 | 1,41 | 0,83 | 1,52 | 0,75 | 1,65 |
| 26 | 1,07 | 1,22 | 1,00 | 1,31 | 0,93 | 1,41 | 0,85 | 1,52 | 0,78 | 1,64 |
| 27 | 1,09 | 1,23 | 1,02 | 1,32 | 0,95 | 1,41 | 0,88 | 1,51 | 0,81 | 1,63 |
| 28 | 1,10 | 1,24 | 1,04 | 1,32 | 0,97 | 1,41 | 0,90 | 1,51 | 0,83 | 1,62 |
| 29 | 1,12 | 1,25 | 1,05 | 1,33 | 0,99 | 1,42 | 0,92 | 1,51 | 0,85 | 1,61 |
| 30 | 1,13 | 1,26 | 1,07 | 1,34 | 1,01 | 1,42 | 0,94 | 1,51 | 0,88 | 1,61 |
| 35 | 1,19 | 1,31 | 1,14 | 1,37 | 1,08 | 1,44 | 1,03 | 1,51 | 0,97 | 1,59 |
| 40 | 1,25 | 1,34 | 1,20 | 1,40 | 1,15 | 1,46 | 1,10 | 1,52 | 1,05 | 1,58 |
| 45 | 1,29 | 1,38 | 1,24 | 1,42 | 1,20 | 1,48 | 1,16 | 1,53 | 1,11 | 1,58 |
| 50 | 1,32 | 1,40 | 1,28 | 1,45 | 1,24 | 1,49 | 1,20 | 1,54 | 1,16 | 1,59 |
| 60 | 1,38 | 1,45 | 1,35 | 1,48 | 1,32 | 1,52 | 1,28 | 1,56 | 1,25 | 1,60 |
| 70 | 1,43 | 1,49 | 1,40 | 1,52 | 1,37 | 1,55 | 1,34 | 1,58 | 1,31 | 1,61 |
| 80 | 1,47 | 1,52 | 1,44 | 1,54 | 1,42 | 1,57 | 1,39 | 1,60 | 1,36 | 1,62 |
| 90 | 1,50 | 1,54 | 1,47 | 1,56 | 1,45 | 1,59 | 1,43 | 1,61 | 1,41 | 1,64 |
| 100 | 1,52 | 1,56 | 1,50 | 1,58 | 1,48 | 1,60 | 1,46 | 1,63 | 1,44 | 1,65 |

# REGRESSION LINEAIRE SIMPLE

## NORMALITE DES RESIDUS

Il faut vérifier l'hypothèse de normalité sur les termes d'erreur. Cette hypothèse n'est pas indispensable, mais elle permet d'utiliser les tests statistiques de ce document qui reposent sur cette hypothèse. Pour cela , on va tester la normalité des résidus par les méthodes disponibles dans **la norme NF X 06-050 (**voir annexe B).

Dans cette norme, il faut retenir essentiellement les méthodes graphiques et statistiques suivantes :

- Fréquences cumulées (droite de Henry)
- Test de Shapiro-Wilk
- Test de Kolmogorov-Smirnov
- Coefficient d'asymétrie
- Coefficient d'aplatissement

# REGRESSION LINEAIRE SIMPLE

## Qualité d'ajustement, tests, intervalles de confiance

*On trouve dans cette fiche l'explication de résultats statistiques classiques souvent fournis par les logiciels.*

### Qualité de l'ajustement

Formule de décomposition, coefficient de détermination

Les variations de *Y* sont expliquées d'une part par les variations de *X*, et d'autre part par le caractère aléatoire des expériences. Pour préciser la part de variation de *Y* expliquée par la variation de *X* et celle aléatoire, on utilise la relation suivante :

**Variation totale = Variation due à la régression + variation résiduelle**

Ce qu'on écrit :

$$SC_{Total} = SC_{Régression} + SC_{Résiduel}$$

ou encore

$$\sum_{i=1}^{n}(Y_i - \overline{Y})^2 = \sum_{i=1}^{n}(\hat{Y}_i - \overline{Y}_i)^2 + \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

(Cette relation est souvent appelée « formule de décomposition de la variance». )

où :

| | | |
|---|---|---|
| $SC_{Total}$ ou « total sum of square » (SST) ou « somme des carrés totale » | mesure la variation globale des $y_i$ autour de leur moyenne $\overline{y}$ . | $\sum_{i=1}^{n}(Y_i - \overline{Y})^2$ |
| $SC_{Régression}$ ou « regression sum of square » (SSR) ou « somme des carrés expliquée par *X* » | mesure la part de variation de *Y* expliquée par le modèle (par une variation de *X*). | $\sum_{i=1}^{n}(\hat{Y}_i - \overline{Y})^2$ |
| $SC_{Résiduel}$ ou « error sum of square » (SSE) ou « somme des carrés résiduelle » | mesure la part de variation de *Y* non expliquée par *X* | $\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$ |

Avec ces notations, on peut alors introduire la notion de coefficient de détermination. Ce coefficient exprime la part de variation de *Y* expliquée par la variation de *X.* Il exprime le rapport entre la variance de *Y* expliquée par le modèle et la variance totale. C'est un indicateur de la qualité de l'ajustement de l'équation de régression. Il permet d'avoir une idée globale de l'ajustement du modèle. Il est noté $R^2$ et est défini par la relation suivante :

$$R^2 = \frac{SC_{Régression}}{SC_{Total}} = 1 - \frac{\sum_{i=1}^{n}(Y_i - \hat{Y})^2_i}{\sum_{i=1}^{n}(Y_i - \overline{Y})^2}$$

Un $R^2$ proche de 1 signifie a priori que l'ajustement est bon mais il ne faut surtout pas se fier uniquement à ce résultat. Il faut aussi étudier les résidus ainsi que l'écart-type résiduel. On ne pourra conclure sur la qualité du modèle qu'après ces démarches.

**Remarques importantes :**

- Un coefficient de détermination $R^2$ égal à 1 correspond à une situation où tous les points observés sont sur la droite de régression puisqu'on a $\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 = 0$. La liaison entre *Y* et *X* est alors parfaitement linéaire et les résidus sont nuls. Cependant, il faut faire très attention car les logiciels peuvent donner un coefficient de détermination $R^2$ égal à 1 sans pour autant donner des résidus nuls. C 'est parce qu'il ne faut pas oublier que les logiciels donnent un arrondi de $R^2$, ils donnent donc 1 si sa valeur est par exemple 0.9999999.

- Un coefficient de détermination $R^2$ égal à 0 correspond à une situation où la droite des moindres carrés est horizontale, puisqu'on a $\sum_{i=1}^{n}(\hat{Y}_i - \overline{Y})^2 = 0$. Il n'y a aucune liaison linéaire entre *Y* et *X*.

- Le coefficient de corrélation entre les variables *X* et *Y (noté R)* élevé au carré est égal au coefficient de détermination. Il mesure à la fois la force et le sens de la liaison entre *X* et *Y* . Un *R* positif signifie que si *X* croit alors *Y* croit, si *X* décroît alors *Y* décroît. Un *R* négatif signifie que si *X* croit alors *Y* décroît, si *X* décroît alors *Y* croit.

**Tests statistiques**

*ATTENTION : Il est impératif d'avoir vérifier l'hypothèse de normalité des résidus (voir fiche « normalité des résidus ») pour utiliser les tests statistiques.*

***Le test de FISHER-SNEDECOR***

**Ce test mesure l'apport global de la variable *X* sur la détermination de *Y*. On l'utilise pour savoir si le modèle est significatif.**

Dans le cas d'une régression linéaire simple, il consiste à tester la significativité de la pente $\beta_1$. Il est alors strictement équivalent au test de Student effectué pour tester la pente $\beta_1$ (en effet on a la relation suivante entre *le F* de Fisher et le *t* de Student : $t^2_{n-2} = F(1, n-2)$ ).

L'hypothèse testée dans le cadre d'une ***régression linéaire simple*** est :

- **H0 :** $\beta_1 = 0$

contre

- **H1 :** $\beta_1 \neq 0$

On appelle *F* la statistique de test de Fisher-Snedecor, elle est donnée pour la régression simple dans le tableau récapitulatif suivant :

| Source de variation | Degré de liberté | Somme des carrés (des écarts) (SC) | Variance | F |
|---|---|---|---|---|
| **Régression** | 1 | $SC_{Régression}$ | $SC_{Régression}$ / 1 | $\frac{Variance_{Régression}}{Variance_{Résiduel}}$ |
| **Résiduel** | n - 2 | $SC_{Résiduel}$ | $SCE_{Résiduel}$ / (n-2) | |
| **Total** | n - 1 | $SC_{Total}$ | | |

La table du test de Fisher-Snedecor fournit une valeur $F_{1-\alpha}(\nu_1, \nu_2)$ en fonction du nombre de degrés de liberté du numérateur $\nu_1$ (dans le cas de la régression simple, $\nu_1$ =1), du nombre de degrés de liberté du numérateur $\nu_2$ (dans le cas de la régression simple, $\nu_2$ =n-2) et de la valeur du risque $\alpha$.

Après avoir fixé la valeur du risque $\alpha$ , on conclut au rejet ou non-rejet de la significativité de $\beta_1$ de la façon suivante :

1. Si la statistique *F* du test (calculée avec les données) est inférieure ou égale à $F_{1-\alpha}(\nu_1, \nu_2)$ l'hypothèse H0 est acceptée, c'est à dire que $\beta_1$ n'est pas significatif.

2. Si la statistique *F* du test est supérieure à $F_{1-\alpha}(\nu_1, \nu_2)$ l'hypothèse H0 est rejetée, c'est à dire que $\beta_1$ est significatif.

## TABLE DE FISHER-SNEDECOR

*Fractiles $F_{0,95}(\nu_1, \nu_2)$ de la loi de Fisher-Snedecor*

| DEGRÉS DE LIBERTÉ DU DÉNOMINATEUR : $\nu_2$ \ DEGRÉS DE LIBERTÉ DU NUMÉRATEUR : $\nu_1$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 12 | 14 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 161 | 200 | 216 | 225 | 230 | 234 | 237 | 239 | 241 | 242 | 244 | 245 |
| 2 | 18,5 | 19,0 | 19,2 | 19,2 | 19,3 | 19,3 | 19,4 | 19,4 | 19,4 | 19,4 | 19,4 | 19,4 |
| 3 | 10,1 | 9,55 | 9,28 | 9,12 | 9,01 | 8,94 | 8,89 | 8,85 | 8,81 | 8,79 | 8,74 | 8,71 |
| 4 | 7,71 | 6,94 | 6,59 | 6,39 | 6,26 | 6,16 | 6,09 | 6,04 | 6,00 | 5,96 | 5,91 | 5,87 |
| 5 | 6,61 | 5,79 | 5,41 | 5,19 | 5,05 | 4,95 | 4,88 | 4,82 | 4,77 | 4,74 | 4,68 | 4,64 |
| 6 | 5,99 | 5,14 | 4,76 | 4,53 | 4,39 | 4,28 | 4,21 | 4,15 | 4,10 | 4,06 | 4,00 | 3,96 |
| 7 | 5,59 | 4,74 | 4,35 | 4,12 | 3,97 | 3,87 | 3,79 | 3,73 | 3,68 | 3,64 | 3,57 | 3,53 |
| 8 | 5,32 | 4,46 | 4,07 | 3,84 | 3,69 | 3,58 | 3,50 | 3,44 | 3,39 | 3,35 | 3,28 | 3,24 |
| 9 | 5,12 | 4,26 | 3,86 | 3,63 | 3,48 | 3,37 | 3,29 | 3,23 | 3,18 | 3,14 | 3,07 | 3,03 |
| 10 | 4,96 | 4,10 | 3,71 | 3,48 | 3,33 | 3,22 | 3,14 | 3,07 | 3,02 | 2,98 | 2,91 | 2,86 |
| 11 | 4,84 | 3,98 | 3,59 | 3,36 | 3,20 | 3,09 | 3,01 | 2,95 | 2,90 | 2,85 | 2,79 | 2,74 |
| 12 | 4,75 | 3,89 | 3,49 | 3,26 | 3,11 | 3,00 | 2,91 | 2,85 | 2,80 | 2,75 | 2,69 | 2,64 |
| 13 | 4,67 | 3,81 | 3,41 | 3,18 | 3,03 | 2,92 | 2,83 | 2,77 | 2,71 | 2,67 | 2,60 | 2,55 |
| 14 | 4,60 | 3,74 | 3,34 | 3,11 | 2,96 | 2,85 | 2,76 | 2,70 | 2,65 | 2,60 | 2,53 | 2,48 |
| 15 | 4,54 | 3,68 | 3,29 | 3,06 | 2,90 | 2,79 | 2,71 | 2,64 | 2,59 | 2,54 | 2,48 | 2,42 |
| 16 | 4,49 | 3,63 | 3,24 | 3,01 | 2,85 | 2,74 | 2,66 | 2,59 | 2,54 | 2,49 | 2,42 | 2,37 |
| 17 | 4,45 | 3,59 | 3,20 | 2,96 | 2,81 | 2,70 | 2,61 | 2,55 | 2,49 | 2,45 | 2,38 | 2,33 |
| 18 | 4,41 | 3,55 | 3,16 | 2,93 | 2,77 | 2,66 | 2,58 | 2,51 | 2,46 | 2,41 | 2,34 | 2,29 |
| 19 | 4,38 | 3,52 | 3,13 | 2,90 | 2,74 | 2,63 | 2,54 | 2,48 | 2,42 | 2,38 | 2,31 | 2,26 |
| 20 | 4,35 | 3,49 | 3,10 | 2,87 | 2,71 | 2,60 | 2,51 | 2,45 | 2,39 | 2,35 | 2,28 | 2,22 |
| 21 | 4,32 | 3,47 | 3,07 | 2,84 | 2,68 | 2,57 | 2,49 | 2,42 | 2,37 | 2,32 | 2,25 | 2,20 |
| 22 | 4,30 | 3,44 | 3,05 | 2,82 | 2,66 | 2,55 | 2,46 | 2,40 | 2,34 | 2,30 | 2,23 | 2,17 |
| 23 | 4,28 | 3,42 | 3,03 | 2,80 | 2,64 | 2,53 | 2,44 | 2,37 | 2,32 | 2,27 | 2,20 | 2,15 |
| 24 | 4,26 | 3,40 | 3,01 | 2,78 | 2,62 | 2,51 | 2,42 | 2,36 | 2,30 | 2,25 | 2,18 | 2,13 |
| 25 | 4,24 | 3,39 | 2,99 | 2,76 | 2,60 | 2,49 | 2,40 | 2,34 | 2,28 | 2,24 | 2,16 | 2,11 |
| 26 | 4,23 | 3,37 | 2,98 | 2,74 | 2,59 | 2,47 | 2,39 | 2,32 | 2,27 | 2,22 | 2,15 | 2,09 |
| 27 | 4,21 | 3,35 | 2,96 | 2,73 | 2,57 | 2,46 | 2,37 | 2,31 | 2,25 | 2,20 | 2,13 | 2,08 |
| 28 | 4,20 | 3,34 | 2,95 | 2,71 | 2,56 | 2,45 | 2,36 | 2,29 | 2,24 | 2,19 | 2,12 | 2,06 |
| 29 | 4,18 | 3,33 | 2,93 | 2,70 | 2,55 | 2,43 | 2,35 | 2,28 | 2,22 | 2,18 | 2,10 | 2,05 |
| 30 | 4,17 | 3,32 | 2,92 | 2,69 | 2,53 | 2,42 | 2,33 | 2,27 | 2,21 | 2,16 | 2,09 | 2,04 |
| 32 | 4,15 | 3,29 | 2,90 | 2,67 | 2,51 | 2,40 | 2,31 | 2,24 | 2,19 | 2,14 | 2,07 | 2,01 |
| 34 | 4,13 | 3,28 | 2,88 | 2,65 | 2,49 | 2,38 | 2,29 | 2,23 | 2,17 | 2,12 | 2,05 | 1,99 |
| 36 | 4,11 | 3,26 | 2,87 | 2,63 | 2,48 | 2,36 | 2,28 | 2,21 | 2,15 | 2,11 | 2,03 | 1,98 |
| 38 | 4,10 | 3,24 | 2,85 | 2,62 | 2,46 | 2,35 | 2,26 | 2,19 | 2,14 | 2,09 | 2,02 | 1,96 |
| 40 | 4,08 | 3,23 | 2,84 | 2,61 | 2,45 | 2,34 | 2,25 | 2,18 | 2,12 | 2,08 | 2,00 | 1,95 |
| 50 | 4,03 | 3,18 | 2,79 | 2,56 | 2,40 | 2,29 | 2,20 | 2,13 | 2,07 | 2,03 | 1,95 | 1,89 |
| 60 | 4,00 | 3,15 | 2,76 | 2,53 | 2,37 | 2,25 | 2,17 | 2,10 | 2,04 | 1,99 | 1,92 | 1,86 |
| 70 | 3,98 | 3,13 | 2,74 | 2,50 | 2,35 | 2,23 | 2,14 | 2,07 | 2,02 | 1,97 | 1,89 | 1,84 |
| 80 | 3,96 | 3,11 | 2,72 | 2,49 | 2,33 | 2,21 | 2,13 | 2,06 | 2,00 | 1,95 | 1,88 | 1,82 |
| 90 | 3,95 | 3,10 | 2,71 | 2,47 | 2,32 | 2,20 | 2,11 | 2,04 | 1,99 | 1,94 | 1,86 | 1,80 |
| 100 | 3,94 | 3,09 | 2,70 | 2,46 | 2,31 | 2,19 | 2,10 | 2,03 | 1,97 | 1,93 | 1,85 | 1,79 |
| ∞ | 3,84 | 3,00 | 2,60 | 2,37 | 2,21 | 2,10 | 2,01 | 1,94 | 1,88 | 1,83 | 1,75 | 1,69 |

### *Tests d'hypothèses sur les coefficients de régression $\beta_0$ et $\beta_1$*

Les logiciels statistiques effectuent systématiquement des tests pour regarder si les coefficients de régression sont significatifs.

#### Test sur $\beta_0$

On veut tester l'hypothèse suivante :

- **H0 :** $\beta_0 = 0$ (droite passant par l'origine)

contre

- **H1 :** $\beta_0 \neq 0$

La statistique utilisée pour ce test est : $$T = \frac{b_0}{\sqrt{\text{var}(b_0)}} = \frac{b_0}{s\sqrt{\frac{1}{n} + \frac{\overline{X}}{\sum_{i=1}^{n}(X_i - \overline{X})^2}}}$$

La loi de probabilité de *T* sous l'hypothèse H0 vraie est une loi de Student à n-2 degrés de liberté.

La table du test de Student fournit une valeur $t_{1-\alpha/2}(\nu)$ en fonction du nombre de degrés de liberté $\nu$ (n-2) et de la valeur du risque $\alpha$.

Après avoir fixé la valeur du risque $\alpha$, on conclut au rejet ou non-rejet de l'hypothèse nulle de la façon suivante :

1. Si la valeur absolue de la statistique *T* du test (calculée avec les données) est inférieure ou égale à $t_{1-\alpha/2}(\nu)$ l'hypothèse H0 est acceptée, c'est à dire que $\beta_0$ n'est pas significativement différent de zéro au risque $\alpha$.
2. Si la valeur absolue de la statistique *T* du test est supérieure à $t_{1-\alpha/2}(\nu)$ l'hypothèse H0 est rejetée, c'est à dire que $\beta_0$ est significativement différent de zéro au risque $\alpha$.

Test sur $\beta_1$

On veut tester l'hypothèse suivante :

- **H0 :** $\beta_1 = 0$

contre

- **H1 :** $\beta_1 \neq 0$

La statistique utilisée pour ce test est :

$$T = \frac{b_1}{\sqrt{\text{var}(b_1)}} = \frac{b_1}{s\sqrt{\dfrac{1}{\sum_{i=1}^{n}(X_i - \bar{X})^2}}}$$

La loi de probabilité de *T* sous l'hypothèse H0 vraie est une loi de Student à n-2 degrés de liberté.

La table du test de Student fournit une valeur $t_{1-\alpha/2}(\nu)$ en fonction du nombre de degrés de liberté $\nu$ (n-2) et de la valeur du risque $\alpha$.

Après avoir fixé la valeur du risque $\alpha$, on conclut au rejet ou non-rejet de l'hypothèse nulle de la façon suivante :

3. Si la valeur absolue de la statistique *T* du test (calculée avec les données) est inférieure ou égale à $t_{1-\alpha/2}(\nu)$ l'hypothèse H0 est acceptée, c'est à dire que $\beta_1$ n'est pas significativement différent de zéro au risque $\alpha$.
4. Si la valeur absolue de la statistique *T* du test est supérieure à $t_{1-\alpha/2}(\nu)$ l'hypothèse H0 est rejetée, c'est à dire que $\beta_1$ est significativement différent de zéro au risque $\alpha$.

Remarque: pour tester l'hypothèse $\beta_1 = 1$, on procède de la même manière en utilisant la statistique : $T = \dfrac{b_1 - 1}{\sqrt{\text{var}(b_1)}}$

*Fractiles de la loi de Student*

Cette table donne les valeurs des fractiles $t_P(\nu)$ de la loi de Student pour $P \geq 0{,}60$. Pour les valeurs $P \leq 0{,}40$, on a $t_P(\nu) = -t_{1-P}(\nu)$.

| $\nu$ \ $P$ | 0,60 | 0,70 | 0,80 | 0,90 | 0,95 | 0,975 | 0,990 | 0,995 | 0,999 | 0,9995 |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 0,325 | 0,727 | 1,376 | 3,078 | 6,314 | 12,71 | 31,82 | 63,66 | 318,3 | 636,6 |
| 2 | 0,289 | 0,617 | 1,061 | 1,886 | 2,920 | 4,303 | 6,965 | 9,925 | 22,33 | 31,60 |
| 3 | 0,277 | 0,584 | 0,978 | 1,638 | 2,353 | 3,182 | 4,541 | 5,841 | 10,22 | 12,94 |
| 4 | 0,271 | 0,569 | 0,941 | 1,533 | 2,132 | 2,776 | 3,747 | 4,604 | 7,173 | 8,610 |
| 5 | 0,267 | 0,559 | 0,920 | 1,476 | 2,015 | 2,571 | 3,365 | 4,032 | 5,893 | 6,859 |
| 6 | 0,265 | 0,553 | 0,906 | 1,440 | 1,943 | 2,447 | 3,143 | 3,707 | 5,208 | 5,959 |
| 7 | 0,263 | 0,549 | 0,896 | 1,415 | 1,895 | 2,365 | 2,998 | 3,499 | 4,785 | 5,405 |
| 8 | 0,262 | 0,546 | 0,889 | 1,397 | 1,860 | 2,306 | 2,896 | 3,355 | 4,501 | 5,041 |
| 9 | 0,261 | 0,543 | 0,883 | 1,383 | 1,833 | 2,262 | 2,821 | 3,250 | 4,297 | 4,781 |
| 10 | 0,260 | 0,542 | 0,879 | 1,372 | 1,812 | 2,228 | 2,764 | 3,169 | 4,144 | 4,587 |
| 11 | 0,260 | 0,540 | 0,876 | 1,363 | 1,796 | 2,201 | 2,718 | 3,106 | 4,025 | 4,437 |
| 12 | 0,259 | 0,539 | 0,873 | 1,356 | 1,782 | 2,179 | 2,681 | 3,055 | 3,930 | 4,318 |
| 13 | 0,259 | 0,538 | 0,870 | 1,350 | 1,771 | 2,160 | 2,650 | 3,012 | 3,852 | 4,221 |
| 14 | 0,258 | 0,537 | 0,868 | 1,345 | 1,761 | 2,145 | 2,624 | 2,977 | 3,787 | 4,140 |
| 15 | 0,258 | 0,536 | 0,866 | 1,341 | 1,753 | 2,131 | 2,602 | 2,947 | 3,733 | 4,073 |
| 16 | 0,258 | 0,535 | 0,865 | 1,337 | 1,746 | 2,120 | 2,583 | 2,921 | 3,686 | 4,015 |
| 17 | 0,257 | 0,534 | 0,863 | 1,333 | 1,740 | 2,110 | 2,567 | 2,898 | 3,646 | 3,965 |
| 18 | 0,257 | 0,534 | 0,862 | 1,330 | 1,734 | 2,101 | 2,552 | 2,878 | 3,611 | 3,922 |
| 19 | 0,257 | 0,533 | 0,861 | 1,328 | 1,729 | 2,093 | 2,539 | 2,861 | 3,579 | 3,883 |
| 20 | 0,257 | 0,533 | 0,860 | 1,325 | 1,725 | 2,086 | 2,528 | 2,845 | 3,552 | 3,850 |
| 21 | 0,257 | 0,532 | 0,859 | 1,323 | 1,721 | 2,080 | 2,518 | 2,831 | 3,527 | 3,819 |
| 22 | 0,256 | 0,532 | 0,858 | 1,321 | 1,717 | 2,074 | 2,508 | 2,819 | 3,505 | 3,792 |
| 23 | 0,256 | 0,532 | 0,858 | 1,319 | 1,714 | 2,069 | 2,500 | 2,807 | 3,485 | 3,767 |
| 24 | 0,256 | 0,531 | 0,857 | 1,318 | 1,711 | 2,064 | 2,492 | 2,797 | 3,467 | 3,745 |
| 25 | 0,256 | 0,531 | 0,856 | 1,316 | 1,708 | 2,060 | 2,485 | 2,787 | 3,450 | 3,725 |
| 26 | 0,256 | 0,531 | 0,856 | 1,315 | 1,706 | 2,056 | 2,479 | 2,779 | 3,435 | 3,707 |
| 27 | 0,256 | 0,531 | 0,855 | 1,314 | 1,703 | 2,052 | 2,473 | 2,771 | 3,421 | 3,690 |
| 28 | 0,256 | 0,530 | 0,855 | 1,313 | 1,701 | 2,048 | 2,467 | 2,763 | 3,408 | 3,674 |
| 29 | 0,256 | 0,530 | 0,854 | 1,311 | 1,699 | 2,045 | 2,462 | 2,756 | 3,396 | 3,659 |
| 30 | 0,256 | 0,530 | 0,854 | 1,310 | 1,697 | 2,042 | 2,457 | 2,750 | 3,385 | 3,646 |
| 32 | 0,256 | 0,530 | 0,853 | 1,309 | 1,694 | 2,037 | 2,449 | 2,738 | 3,365 | 3,622 |
| 34 | 0,255 | 0,529 | 0,852 | 1,307 | 1,691 | 2,032 | 2,441 | 2,728 | 3,348 | 3,601 |
| 36 | 0,255 | 0,529 | 0,852 | 1,306 | 1,688 | 2,028 | 2,434 | 2,719 | 3,333 | 3,582 |
| 38 | 0,255 | 0,529 | 0,851 | 1,304 | 1,686 | 2,024 | 2,429 | 2,712 | 3,319 | 3,566 |
| 40 | 0,255 | 0,529 | 0,851 | 1,303 | 1,684 | 2,021 | 2,423 | 2,704 | 3,307 | 3,551 |
| 50 | 0,255 | 0,528 | 0,849 | 1,298 | 1,676 | 2,009 | 2,403 | 2,678 | 3,261 | 3,496 |
| 60 | 0,254 | 0,527 | 0,848 | 1,296 | 1,671 | 2,000 | 2,390 | 2,660 | 3,232 | 3,460 |
| 70 | 0,254 | 0,527 | 0,847 | 1,294 | 1,667 | 1,994 | 2,381 | 2,648 | 3,211 | 3,435 |
| 80 | 0,254 | 0,527 | 0,846 | 1,292 | 1,664 | 1,990 | 2,374 | 2,639 | 3,195 | 3,415 |
| 90 | 0,254 | 0,526 | 0,846 | 1,291 | 1,662 | 1,987 | 2,368 | 2,632 | 3,183 | 3,402 |
| 100 | 0,254 | 0,526 | 0,845 | 1,290 | 1,660 | 1,984 | 2,365 | 2,626 | 3,174 | 3,389 |
| 200 | 0,254 | 0,525 | 0,843 | 1,286 | 1,653 | 1,972 | 2,345 | 2,601 | 3,131 | 3,339 |
| 500 | 0,253 | 0,525 | 0,842 | 1,283 | 1,648 | 1,965 | 2,334 | 2,586 | 3,106 | 3,310 |
| $\infty$ | 0,253 | 0,524 | 0,842 | 1,282 | 1,645 | 1,960 | 2,326 | 2,576 | 3,090 | 3,291 |

**Remarque importante sur les tests** :

En général, pour conclure au rejet ou non rejet d'une hypothèse, les logiciels fournissent la **p-valeur**. C'est la probabilité que l'estimation du paramètre testé, prenne la valeur quantitative observée, sous l'hypothèse H0. ***Si la p-value est petite, on rejette H0***.

Concrètement, si l'expérimentateur s'est fixé un seuil de signification $\alpha$ (par exemple 5%), il conclut au rejet ou non rejet de l'hypothèse H0 de la manière suivante :

Rejet de H0 $\Leftrightarrow$ p-value < $\alpha$
Non rejet de H0 $\Leftrightarrow$ p-value > $\alpha$

**Intervalles de confiance**

Il est intéressant de donner les intervalles de confiance de niveau de confiance (1-$\alpha$), des paramètres $\beta_0$, $\beta_1$, $\hat{Y}$, $\hat{\hat{Y}}$. Ces intervalles sont des intervalles aléatoires : ils changent selon le jeu de données.
Les intervalles de confiance suscitent souvent des erreurs d'interprétation et des abus de langage. On entend souvent dire que si I est un intervalle de confiance à 95% d'un paramètre $\theta$, c'est qu'il y a 95% de chance que $\theta$ se trouve dans l'intervalle  Ceci est incorrect. En fait, si on recommence 100 fois notre expérience, qu'on se retrouve avec 100 jeu de données, et qu'on calcule pour chaque jeu de données l'intervalle de confiance, on aura alors 100 intervalles de confiance différents mais a priori tous proches les un des autres. En moyenne, $\theta$ sera dans 95 de ces intervalles.

## Intervalles de confiance de $\beta_0$ et $\beta_1$

Les estimations de $Var(b_0)$ et $Var(b_1)$ sont :

$$Var(b_0) = s^2\left(\frac{1}{n} + \frac{\overline{X}^2}{\sum_{i=1}^{n}(X_i - \overline{X})^2}\right) \qquad Var(b_1) = \frac{s^2}{\sum_{i=1}^{n}(X_i - \overline{X})^2}$$

et les variables $\frac{b_0 - \beta_0}{\sqrt{Var(b_0)}}$ et $\frac{b_1 - \beta_1}{Var(b_1)}$ suivent une loi de Student à n-2 degré de liberté.

D'où les intervalles de confiances bilatéraux suivants où *t* désigne la variable de Student et $1-\alpha$ le niveau de confiance :

$$b_0 - t_{1-\alpha/2}(n-2)\cdot\sqrt{Var(b_0)} \le \beta_0 \le b_0 + t_{1-\alpha/2}(n-2)\cdot\sqrt{Var(b_0)}$$

$$b_1 - t_{1-\alpha/2}(n-2)\cdot\sqrt{Var(b_1)} \le \beta_1 \le b_1 + t_{1-\alpha/2}(n-2)\cdot\sqrt{Var(b_1)}$$

# Intervalles de confiance de $\hat{Y}$ et $\hat{\hat{Y}}$

Rappelons qu'il faut bien différencier les intervalles de confiance de $\hat{Y}$ et $\hat{\hat{Y}}$ malgré l'égalité de leurs estimateurs :

- L'intervalle de confiance de $\hat{Y}$ définit « les limites dans lesquelles se situe probablement une valeur individuelle lue sur la droite de régression : lorsqu'on a construit un modèle qui se présente sous la forme d'une droite de régression, l'intervalle de confiance en question dit que, pour une valeur donnée de la variable *X*, la vraie valeur de la variable *Y* ($\beta_0 + \beta_1 X$) devrait se situer au sein de cet intervalle de confiance. ».
  Nous appellerons cet intervalle : «l'intervalle de confiance d'une prédiction moyenne »
  Avec un niveau de confiance (1-$\alpha$), on l'interprète comme un intervalle dans lequel on a (1-$\alpha$)% de chances de trouver la moyenne d'un très grand nombre d'essais.

- L'intervalle de confiance de $\hat{\hat{Y}}$ définit « les limites dans lesquelles tombera une nouvelle observation de *Y* si elle fait partie de la même population statistique que l'échantillon. Cela produit un intervalle plus large que le précédent : en effet, en plus de la variance de l'échantillon qui a servi à établir l'équation, s'ajoute celle qui est associée au tirage d'un nouvel élément. »

  Nous appellerons cet intervalle : « l'intervalle de confiance d'une prédiction nouvelle »
  Avec un niveau de confiance (1-$\alpha$), il représente l'intervalle dans lequel on peut s'attendre à trouver (1-$\alpha$)% des observations si l'on fait des essais en X= $X_0$

---

- Intervalle de confiance d'une prédiction moyenne pour $X = X_0$

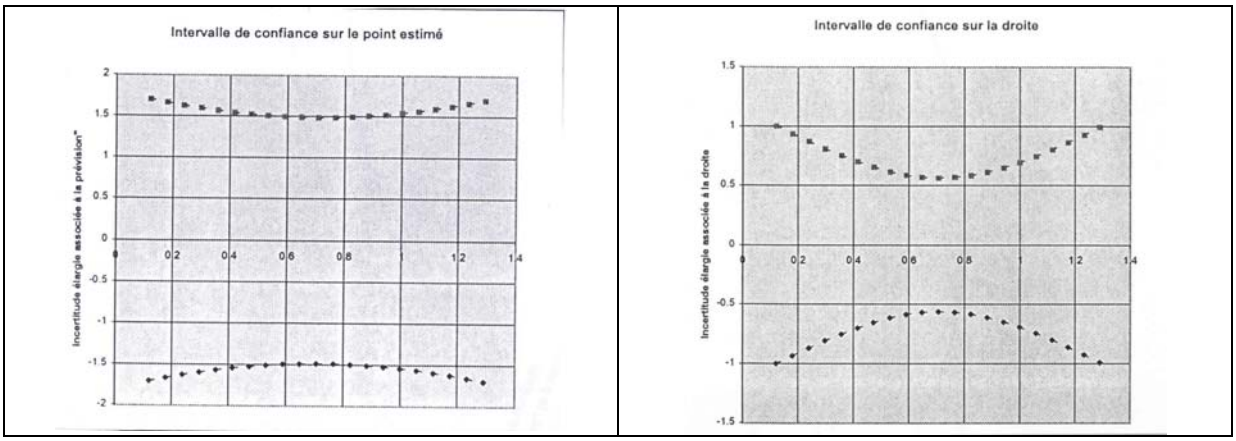
$$b_0 + b_1 X_0 \pm t_{1-\alpha/2}(n-2)s\sqrt{\frac{1}{n} + \frac{(X_0 - \overline{X})^2}{\sum_{i=1}^{n}(X_i - \overline{X})^2}}$$

- Intervalle de confiance d'une prédiction nouvelle pour $X = X_0$

$$b_0 + b_1 X_0 \pm t_{1-\alpha/2}(n-2)s\sqrt{1 + \frac{1}{n} + \frac{(X_0 - \overline{X})^2}{\sum_{i=1}^{n}(X_i - \overline{X})^2}}$$

---

Si l'on calcule les intervalles de confiance de prédiction moyenne ou nouvelle pour toutes les valeurs possibles de *X*, on trouve une région de confiance au niveau (1-$\alpha$) délimitée par deux branches d'hyperbole .



# REGRESSION LINEAIRE SIMPLE

## REGRESSION PONDEREE

Nous avons vu dans les hypothèses que la condition d'homoscédasticité (variance des erreurs constante) doit être vérifiée pour modéliser un phénomène par une régression linéaire simple. Si ce n'est pas le cas, deux solutions sont envisageables : on peut faire un changement de variable sur *Y*, ou effectuer une régression pondérée (weighted least squares).

La régression pondérée prend en compte la différence sur les variances des résidus. Supposons que pour chaque observation i, nous puissions connaître un poids $w_i$, alors estimer les paramètres $\beta_0$ et $\beta_1$ du modèle $Y = \beta_0 + \beta_1 X + \varepsilon$ avec ces pondérations consiste à minimiser la quantité:

$$\sum_{i=1}^{n} w_i (Y_i - \beta_0 - \beta_1 X_i)^2 \text{ ou } \sum_{i=1}^{n} w_i (Y_i - b_0 - b_1 X_i)^2$$

Les estimateurs de $\beta_0$ et $\beta_1$ notés respectivement $b_0$ et $b_1$ sont alors donnés par les formules :

$$\overline{X} = \frac{\sum_{i=1}^{n} w_i X_i}{\sum_{i=1}^{n} w_i}$$

$$\overline{Y} = \frac{\sum_{i=1}^{n} w_i Y_i}{\sum_{i=1}^{n} w_i}$$

$$b_1 = \frac{\sum_{i=1}^{n} w_i (Y_i - \overline{Y})(X_i - \overline{X})}{\sum_{i=1}^{n} w_i (X_i - \overline{X})^2}$$

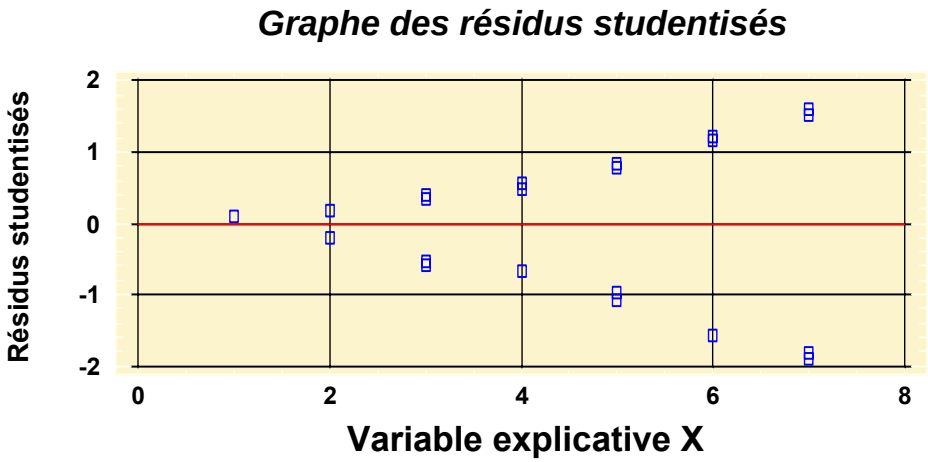
$$b_0 = \overline{Y} - b_1 \overline{X}$$

Les poids permettent de tenir compte de la précision de chaque observation ou de privilégier certains points. On peut les choisir de façon à limiter les influences et à rétablir l'homoscédasticité. De façon générale, on va choisir les poids de la façon suivante :

$$w_i = \frac{\sigma^2}{Var(\varepsilon_i)}$$

Il faut donc évaluer la forme de $Var(\varepsilon_i)$ en fonction de $X_i$ pour connaître les poids à utiliser. Pour déterminer cette forme, on trace généralement le graphe des résidus ou de la variance des résidus en fonction de la variable explicative *X*. On peut alors faire des suppositions sur la forme de $Var(\varepsilon_i)$ en fonction des $X_i$.

Exemple : on obtient les nuage de points suivants:




Le graphe des résidus montre que les écart au modèle choisi ne sont pas constants. Plus *X* augmente, plus l'écart augmente. Il faut donc procéder à une régression pondérée.
Les données recueillies par l'expérimentateur n'interviennent pas avec la même influence dans le calcul du modèle.

On va donc par exemple supposer que :

$$Var(\varepsilon_i) = \sigma^2 X_i^{1/2}$$

ou $Var(\varepsilon_i) = \sigma^2 X_i$

ou $Var(\varepsilon_i) = \sigma^2 X_i^2$

On utilise donc les poids suivants :

$$w_i = \frac{1}{X_i^{1/2}}$$

$$\text{ou } w_i = \frac{1}{X_i}$$

$$\text{ou } w_i = \frac{1}{X_i^2}$$

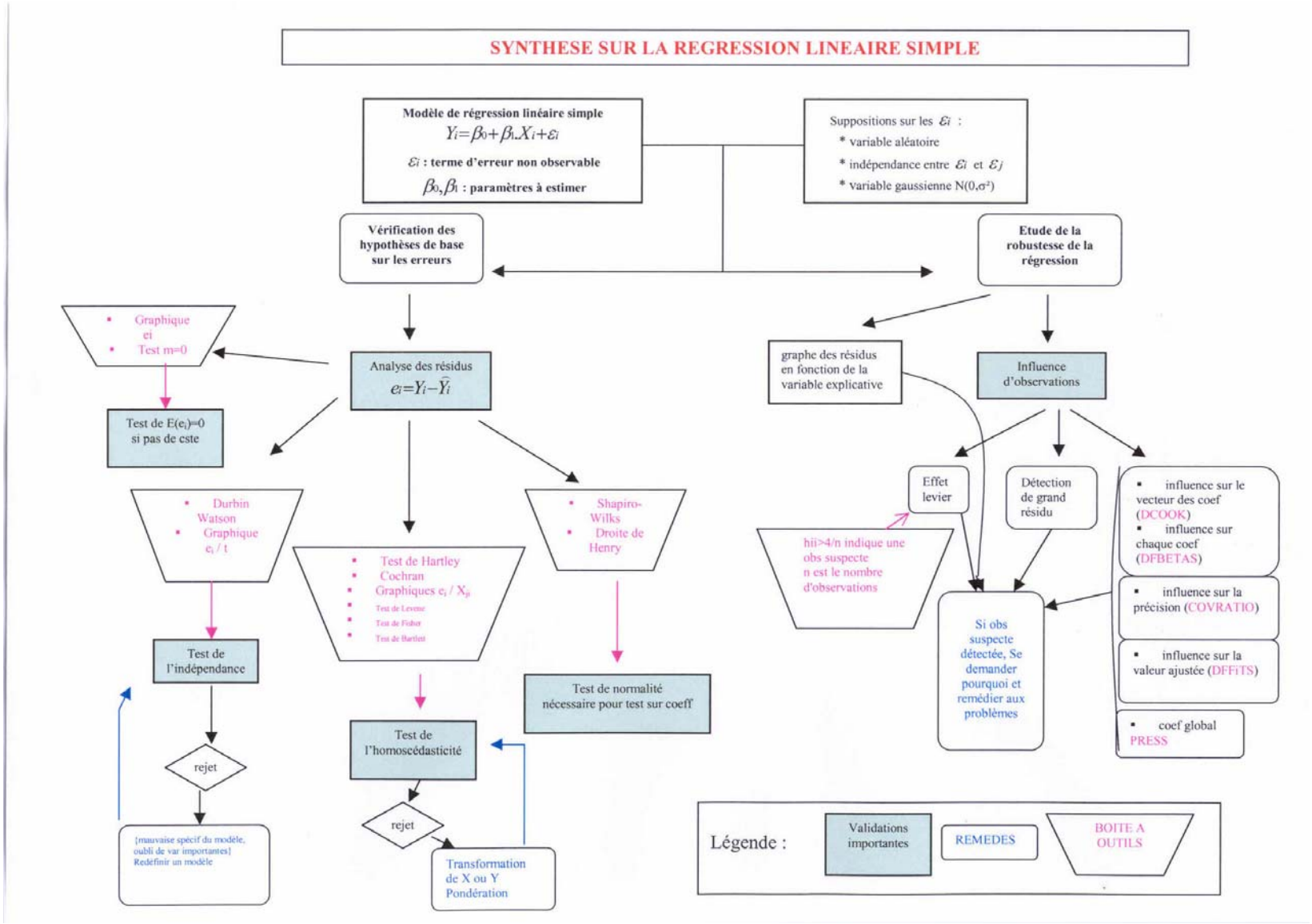
Pour savoir si l'une des formes est meilleure que l'autre, on va effectuer la régression pondérée avec chacune des formes proposées et choisir celle pour laquelle le graphe des résidus en fonction de *X* est le meilleur.

Remarque : dans le cas d'une régression pondérée les résidus ont pour valeur :

$$e_i = w_i (Y_i - b_0 - b_1 X_i)$$

On peut aussi prendre comme poids l'inverse des variances :

$$w_i = \frac{1}{VAR(X_i)}$$


SYNTHESE SUR LA REGRESSION LINEAIRE SIMPLE
Modèle de régression linéaire simple
$Y_i=\beta_0+\beta_1 X_i+\varepsilon_i$
$\varepsilon_i$ : terme d'erreur non observable
$\beta_0,\beta_1$ : paramètres à estimer
Suppositions sur les $\varepsilon_i$ :
* variable aléatoire
* indépendance entre $\varepsilon_i$ et $\varepsilon_j$
* variable gaussienne N(0,$\sigma^2$)
Vérification des hypothèses de base sur les erreurs
Etude de la robustesse de la régression
Graphique ei
Test m=0
Analyse des résidus
$e_i=Y_i-\hat{Y}_i$
graphe des résidus en fonction de la variable explicative
Influence d'observations
Test de $E(e_i)=0$ si pas de cste
Durbin Watson
Graphique $e_i$ / t
Test de Hartley
Cochran
Graphiques $e_i$ / $X_j$
Test de Levene
Test de Fisher
Test de Bartlett
Shapiro-Wilks
Droite de Henry
Effet levier
Détection de grand résidu
hii>4/n indique une obs suspecte
n est le nombre d'observations
influence sur le vecteur des coef (DCOOK)
influence sur chaque coef (DFBETAS)
influence sur la précision (COVRATIO)
influence sur la valeur ajustée (DFFiTS)
coef global PRESS
Si obs suspecte détectée, Se demander pourquoi et remédier aux problèmes
Test de l'indépendance
Test de normalité nécessaire pour test sur coeff
Test de l'homoscédasticité
rejet
rejet
{mauvaise spécif du modèle, oubli de var importantes}
Redéfinir un modèle
Transformation de X ou Y
Pondération
Légende :
Validations importantes
REMEDES
BOITE A OUTILS

# REGRESSION LINEAIRE MULTIPLE

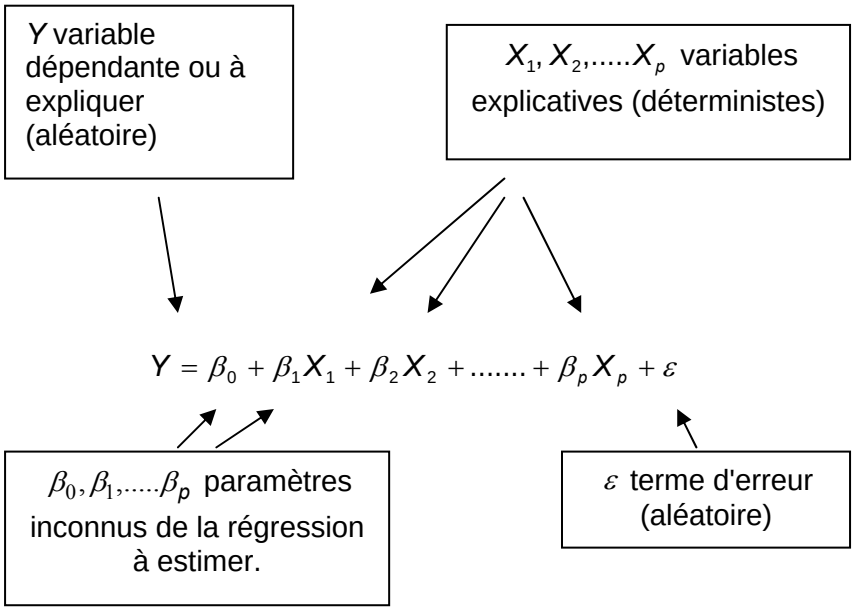
# REGRESSION LINEAIRE MULTIPLE

## LE MODELE ET SES HYPOTHESES.

Il arrive souvent qu'on explique une variable, non pas par une unique variable, mais par un certain nombre de variables. Par exemple, le rendement d'une réaction chimique s'explique par la température et la quantité de catalyseur utilisées pour effectuer la réaction. On parle alors de régression multiple.

La régression multiple étudie la liaison entre une variable dépendante Y, et un ensemble de variables indépendantes $X_1, X_2, .....X_p$. Elle est une généralisation immédiate de la régression simple.

**Le modèle de régression linéaire multiple s'écrit :**



Appliqué à un échantillon de taille n, le modèle de régression multiple devient :

$Y_i$ variable aléatoire à expliquer décrivant les valeurs prises par Y lorsque
$X_1 = X_1^i, X_2 = X_2^i, \ldots\ldots\ldots X_p = X_p^i$

$X_k^i$ ième valeur choisie pour la kième variable explicative $X_k$

$$Y_i = \beta_0 + \beta_1 X_1^i + \beta_2 X_2^i + \ldots\ldots\ldots + \beta_p X_p^i + \varepsilon_i$$

$\beta_0, \beta_1, \ldots\ldots \beta_p$ paramètres inconnus de la régression à estimer.

$\varepsilon_i$ variable aléatoire décrivant les termes d'erreur lorsque :
$X_1 = X_1^i, X_2 = X_2^i, \ldots\ldots\ldots X_p = X_p^i$

Enfin on écrit souvent le modèle de régression multiple sous forme matricielle afin d'exprimer les estimateurs par des formules plus simples.
Pour un échantillon de taille *n* des variables $Y, X_1, X_2, \ldots\ldots\ldots X_p$ , on a :

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

avec :

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ . \\ . \\ . \\ Y_n \end{pmatrix}, \mathbf{X} = \begin{pmatrix} 1 & X_1^1 & X_2^1 & \cdots & \cdots & X_p^1 \\ 1 & X_1^2 & X_2^2 & \cdots & \cdots & X_p^2 \\ 1 & & & & & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & & & & & \vdots \\ 1 & X_1^n & X_2^n & \cdots & \cdots & X_p^n \end{pmatrix}, \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \\ \\ \beta_p \end{pmatrix}, \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \\ \\ \varepsilon_n \end{pmatrix},$$

On utilise des caractères gras pour l'écriture matricielle.

Les hypothèses qui accompagnent le modèle de régression linéaire multiple sont les suivantes :

**HYPOTHESE 1** :Le modèle doit être "linéaire par rapport aux paramètres", c'est à dire que la variable expliquée *Y* s'écrit comme la somme d'une constante (souvent notée $\beta_0$ ,

éventuellement nulle) et de paramètres multipliés par des fonctions des variables explicatives $X_i$.

***Exemples*** :

Rappelons que *p* est le nombre de variables explicatives (la constante n'est pas prise en compte dans *p*)

- $Y = \beta_0 + \beta_1 X_1 + \beta_2 \frac{1}{X_2}$ est un modèle linéaire multiple ($p = 2$)
- $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2^3 + \beta_3 \frac{1}{X_3}$ est un modèle linéaire multiple ($p = 3$)
- $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_1^2 + \beta_3 X_1^3 + \beta_4 X_1^4$ est un modèle linéaire multiple ($p = 4$). malgré qu'il n'y ait pourtant qu'une unique variable explicative. Ce type de modèle est appelé modèle polynomial. C'est un cas particulier de la régression multiple très important c'est pourquoi un chapitre entier lui est souvent réservé dans la littérature. (Voir ANNEXE D)
- $Y = \beta_0 + \beta_1 X_1 + \frac{\beta_2}{\beta_3 + X_2}$ n'est pas un modèle linéaire multiple.

**HYPOTHESE 2** : Le modèle doit être bien spécifié, c'est à dire qu'il ne faut pas avoir oublié de variables explicatives.

**HYPOTHESE 3** : Les variables $X_1, X_2, X_3, \ldots\ldots\ldots X_p$ sont déterministes (ou certaines), c'est à dire que les valeurs qu'elles prennent sont parfaitement connues (contrairement à une variable aléatoire). On parle alors de modèle à effets fixes ou encore de modèle à facteurs contrôlés. (Il existe aussi des modèles à effets aléatoires pour lesquels les valeurs de *X* sont aléatoires.)

**HYPOTHESE 4 : L'espérance des erreurs** associées à chaque p-uplet d'observations est **nulle** :

$$E(\varepsilon_i) = 0 \qquad \forall i \in \{1, n\}$$

Cela implique que la moyenne de $Y_i$ lorsque $X_1 = X_1^i, X_2 = X_2^i, \ldots\ldots\ldots X_p = X_p^i$ est $\beta_0 + \beta_1 X_1^i + \beta_2 X_2^i + \ldots\ldots + \beta_p X_p^i$. Plus généralement, on a :

$$\boxed{E(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots\ldots\ldots + \beta_p X_p}$$

On dit alors que l'estimateur des moindres carrés est **sans biais.**

**HYPOTHESE 5** : La **variance des erreurs** associées à chaque p-uplet d'observations est **constante**. On a donc :

$$Var(\varepsilon_i) = \text{ constante } = \sigma^2 \qquad \forall i \in \{1, n\}$$

L'écart-type des termes d'erreurs est donc constant.
Lorsque cette hypothèse est vérifiée, on dit qu'il y a « **homoscédasticité** »

**HYPOTHESE 6** : Les termes d'erreur ne sont pas corrélés. Il ne doit y avoir aucun lien entre l'erreur commise sur une mesure et les mesures effectuées précédemment.
mathématiquement, cela signifie que :

$$Cov(\varepsilon_i, \varepsilon_j) = 0 \quad pour\ i \neq j$$

**HYPOTHESE 7** : La matrice $\mathbf{X} = \begin{pmatrix} 1 & X_1^1 & X_2^1 & \cdots & \cdots & X_p^1 \\ 1 & X_1^2 & X_2^2 & \cdots & \cdots & X_p^2 \\ 1 & & & & & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & & & & & \vdots \\ 1 & X_1^n & X_2^n & \cdots & \cdots & X_p^n \end{pmatrix}$ doit être de plein rang, de façon à pouvoir inverser X'X où X' est la transposée de X. Concrètement, il faut que les colonnes de cette matrice ne soient pas multiples les unes des autres, c'est à dire mathématiquement que les vecteurs :

$$\mathbf{X_1} = \begin{pmatrix} X_1^1 \\ X_1^2 \\ \vdots \\ \vdots \\ X_1^n \end{pmatrix}, \mathbf{X_2} = \begin{pmatrix} X_2^1 \\ X_2^2 \\ \vdots \\ \vdots \\ X_2^n \end{pmatrix}, \ldots\ldots\ldots\ldots, \mathbf{X_p} = \begin{pmatrix} X_p^1 \\ X_p^2 \\ \vdots \\ \vdots \\ X_p^n \end{pmatrix}$$

ne doivent pas être colinéaires : les variables $X_1, X_2, \ldots\ldots, X_p$ doivent être linéairement indépendantes. Notons que cette hypothèse est propre à la régression multiple et qu'il nous faudra donc préciser plus tard comment la tester.

**HYPOTHESE 8** : **FACULTATIVE** : Les variables aléatoires d'erreur $\varepsilon_i$ suivent une loi normale $N(0, \sigma^2)$.

Cette hypothèse est nécessaire pour l'utilisation des tests statistiques (Student, Fisher …) et la construction d'intervalles de confiance.

Remarque : les variables $Y_i$ suivent alors toutes une loi normale

## LES ESTIMATEURS

L'objectif de la régression est d'estimer les coefficients du modèle ( $\beta_0, \beta_1, \ldots\ldots\beta_p$ ) ainsi que la variance $\sigma^2$ de l'erreur $\varepsilon$ à partir d'un échantillon.

- On appelle $\mathbf{B} = \begin{pmatrix} b_0 \\ b_1 \\ \\ \\ b_p \end{pmatrix}$ l'estimateur de $\boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \\ \\ \beta_p \end{pmatrix}$.

- On appelle $s^2$ l'estimateur de $\sigma^2$.

Ainsi, le modèle de régression "réel" $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta}$, inconnu, est estimé par le modèle $\hat{\mathbf{Y}} = \mathbf{XB}$.

Dans ce document, nous utilisons toujours l'échantillon théorique afin d'obtenir les estimateurs et leurs propriétés. Mais de façon concrète, l'expérimentateur a besoin des estimations. Pour les obtenir, il lui suffit, dans les formules des estimateurs, de remplacer les variables par les données qu'il a obtenu. En toute rigueur, il faudrait différencier les notations utilisées pour les estimateurs et celles pour les estimations, de même il faudrait différencier les notations utilisées pour les données et celles pour les variables aléatoires dont elles sont les réalisations, mais cela complique et alourdit le document, c'est pourquoi nous garderons les mêmes notations, comme dans la plupart des ouvrages sur la régression.

Il existe différentes méthodes pour calculer les estimateurs des coefficients. Nous exposons ici la méthode des moindres carrés. On note $\hat{Y}_i = b_0 + b_1 X_1^i + b_2 X_2^i + ..... + b_p X_p^i$ l'estimateur de la valeur moyenne prise par la variable $Y$ quand $X_1 = X_1^i, X_2 = X_2^i, ..... X_p = X_p^i$. Alors la méthode des moindres carrés consiste à minimiser la somme des carrés des écarts $\hat{Y}_i - Y_i$ , c'est à dire qu'on cherche :

$$\underset{b_0, b_1}{Min} \sum_{i=1}^{n} (Y_i - b_0 - b_1 X_1^i - ..... - b_p X_p^i)^2 = \|\mathbf{Y} - \mathbf{XB}\|^2$$

Ces écarts $Y_i - \hat{Y}_i$ sont appelés les résidus, ce sont des variables aléatoires que nous noterons $e_i$. Ils ne sont pas égaux aux termes d'erreur et contrairement à ces derniers, on peut connaître leur valeur pour chaque échantillon de données.

$\Rightarrow$ On note $e_i = Y_i - \hat{Y}_i$ les résidus

La méthode des moindres carrés conduit aux estimateurs suivants :

| | **ESTIMATEUR** | **MATRICE DE COVARIANCE DE L'ESTIMATEUR (estimée)** |
|---|---|---|
| $\boldsymbol{\beta}$ | $\mathbf{B} = (\mathbf{X'X})^{-1}(\mathbf{X'Y})$ | $COV(\mathbf{B}) = s^2(\mathbf{X'X})^{-1}$ |
| $\sigma^2$ | $\frac{\|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|}{n-p-1} = \frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{n-p-1}$ | |

A NOTER :

- $\mathbf{(X'X)}$ est appelée la matrice d'information
- $\mathbf{(X'X)}^{-1}$ est appelée la matrice de dispersion.

**Il est important de remarquer que *Y* n'intervient pas dans la qualité des estimations contrairement à la matrice d'expériences X.**

## Vérification des hypothèses

La validité de la plupart des résultats que nous exposons dépend directement des hypothèses énoncées auparavant. Il est donc primordial, avant le choix d'un modèle, de réfléchir à chacune des hypothèses et d'utiliser, après l'estimation des paramètres du modèle, des méthodes adéquates pour les valider. Ces méthodes étant les mêmes que pour la régression simple, nous ne les détaillons pas dans cette fiche et vous invitons à lire les fiches suivantes :

- "RESIDUS ET INFLUENCES"
- "NORMALITE DES RESIDUS"
- "HOMOSCEDASTICITE ou égalité des variances"
- "CORRELATION DES TERMES D'ERREUR"

Une seule hypothèse est spécifique à la régression multiple : les variables $X_1, X_2, .....X_p$ doivent être linéairement indépendantes (la matrice d'information doit être inversible).
Les problèmes de multicollinéarité entre les variables peuvent être détectés par l'analyse de la matrice d'expérience et de la matrice de corrélation, mais aussi par le facteur d'inflation de la variance (VIF).

Le VIF est défini par :

$$VIF(X_k) = \frac{1}{1 - R_k^2}$$

où $R_k^2$ est le coefficient de détermination obtenue par la régression de $X_k$ en fonction des (p-1) autres variables. On considère que si $Max(VIF(X_1), VIF(X_2), .....VIF(X_p))$ >10, il y a un problème de multicollinéarité qui va réduire la qualité des estimations.

## LES RESIDUS ET INFLUENCES

On procède de la même manière qu'en régression linéaire simple pour faire l'analyse des résidus, des valeurs atypiques et des points influents.

Voici un tableau récapitulatif des critères à étudier dans le cadre de la régression multiple :

| *Critères de détection de valeurs atypiques* | *Formule* | *Seuil critique à partir duquel l'observation est considérée suspecte* |
|---|---|---|
| **\|Résidu studentisé \| ou \|Rstudent\|** | $es_{(i)} = \frac{e_i}{s_{(i)}\sqrt{1-h_{ii}}}$ | >2 |
| **\|Levier\|** | $h_{ii} = (X(X'X)^{-1}X')_{ii}$ | $> \frac{2(p+1)}{n}$ |
| **\|Distance de Cook\|** | $D_i = \frac{r_i^2 h_{ii}}{(p+1)(1-h_{ii})}$ | $> \frac{4}{(n-p-1)}$ |
| **\|DFITS\|** | $DFITS_i = \frac{\hat{Y}_i - \hat{Y}_{(i)}}{s_{(i)}\sqrt{h_{ii}}}$ | $> 2\sqrt{\frac{(p+1)}{n}}$ |
| **\|DFBETAS\|** | $dfbetas_{i,j} = \frac{b_j - b_{j(i)}}{s_{(i)}\sqrt{(X'X)^{-1}_{jj}}}$ | $> \frac{2}{\sqrt{n}}$ |
| **\|COVRATIO\|** | $Covratio_i = \frac{s_{(i)}^2 \det(X'X)^{-1}_{(i)}}{s_{(i)}^2 \det(X'X)^{-1}}$ | >1 augmente la qualité de l'estimation<br><1 diminue la qualité de l'estimation |

## LES PREDICTIONS

On souhaite prédire la valeur de *Y* d'une nouvelle observation pour laquelle $X_1 = X_1^k, X_2 = X_2^k, \ldots\ldots\ldots X_p = X_p^k$. On note cette observation sous forme matricielle $\mathbf{X}^k = \begin{pmatrix} 1 & X_1^k & X_2^k & \cdots & X_p^k \end{pmatrix}$ où $\mathbf{X}^k$ est une matrice (1,(*p*+1))Il faut alors distinguer deux sortes de prédictions :

1. $\hat{Y}_k$ la prédiction sur la moyenne de *Y* pour la nouvelle observation $\mathbf{X}^k$. C'est l'estimateur de la valeur moyenne prise par la variable *Y* quand. $X_1 = X_1^k, X_2 = X_2^k, \ldots\ldots\ldots X_p = X_p^k$.
2. $\hat{\hat{Y}}_k$ la prédiction sur une nouvelle valeur de *Y* pour la nouvelle observation $\mathbf{X}^k$ (résultat d'un nouvel essai au point $\mathbf{X}^k = \begin{pmatrix} X_1^k & X_2^k & \cdots & X_p^k \end{pmatrix}$). C'est l'estimateur de la valeur que prendra la variable *Y* quand $X_1 = X_1^k, X_2 = X_2^k, \ldots\ldots\ldots X_p = X_p^k$.

Les estimateurs $\hat{Y}_k$ et $\hat{\hat{Y}}_k$ sont égaux, mais ils n'estiment pas la même chose et n'ont d'ailleurs pas la même variance (c'est pourquoi il est important de bien différencier les deux).

| | **ESTIMATEUR** | **VARIANCE DE L'ESTIMATEUR** |
|---|---|---|
| $\mathbf{Y/X}^k$ $(= \beta_0 + \beta_1 X_1^k + \ldots\ldots + \beta_p X_p^k)$ | $\hat{Y}_k = b_0 + b_1 X_1^k + \ldots. + b_p X_p^k$ $= \mathbf{X^k B}$ | $1 + (\mathbf{X}^k)(\mathbf{X'X})^{-1}(\mathbf{X}^k)'$ |
| $E(\mathbf{Y/X}^k)$ $(= \beta_0 + \beta_1 X_1^k + \ldots\ldots + \beta_p X_p^k + \varepsilon_k)$ | $\hat{Y}_k = b_0 + b_1 X_1^k + \ldots. + b_p X_p^k$ $= \mathbf{X^k B}$ | $(\mathbf{X}^k)(\mathbf{X'X})^{-1}(\mathbf{X}^k)'$ |

**A NOTER** : l'annexe A est valable pour la régression multiple, elle introduit la "hat matrice" H qui est utile pour écrire le modèle prédictif sous forme matricielle. En effet, on a :

$$\begin{aligned} \hat{\mathbf{Y}} &= \mathbf{XB} \\ &= \mathbf{(X)(X'X)^{-1}(X'Y)} \\ &= \mathbf{HY} \end{aligned}$$

Où **X'** est la transposée de **X** et **H** est la «matrice chapeau » ( « hat matrice »).

## Qualité de l'ajustement

Formule de décomposition, coefficient de détermination

La formule de décomposition présentée en régression simple reste vraie en régression multiple :

$$SC_{Total} = SC_{Régression} + SC_{Résiduel}$$

ou

$$\sum_{i=1}^{n}(Y_i - \overline{Y})^2 = \sum_{i=1}^{n}(\hat{Y}_i - \overline{Y}_i)^2 + \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

(Voir fiche "QUALITE D'AJUSTEMENT")

De même, la formule donnant le coefficient de détermination $R^2$ reste la même :

$$R^2 = \frac{SC_{Régression}}{SC_{Total}} = 1 - \frac{\sum_{i=1}^{n}(Y_i - \hat{Y})^2{}_i}{\sum_{i=1}^{n}(Y_i - \overline{Y})^2}$$

Cependant, le coefficient de détermination pose un problème : il a la propriété d'être croissant en fonction du nombre de variables. Si on a un grand nombre de variables, il aura donc tendance à se rapprocher de 1 sans pour autant que le modèle soit bon. De plus si on veut comparer deux modélisations pour un même phénomène, il faudra que les deux modèles ait le même nombre de variables sinon leur $R^2$ ne seront pas comparables.

C'est pour remédier à ce problème que le $R^2$ ajusté a été introduit. C'est un coefficient de détermination qui introduit une pénalisation liée au nombre de variables. Sa formule est :

$$R^2 ajusté = 1 - \frac{\sum_{i=1}^{n}\frac{(Y_i - \hat{Y}_i)^2}{n - p - 1}}{\sum_{i=1}^{n}\frac{(Y_i - \overline{Y})^2}{n-1}}$$

Pour résumer, le $R^2$ ajusté autorise la comparaison de plusieurs équations de régression multiple comportant la même variable dépendante mais dont les équations diffèrent par le nombre de variables explicatives *p*.
C'est un critère d'une grande utilité pour la recherche du modèle optimal.

## ***Les tests d'hypothèses***

***Le test de FISHER-SNEDECOR***

Ce test mesure l'apport global des variables explicatives sur la détermination de *Y*. Il teste si la liaison globale entre *Y* et $X_1, X_2, .....X_p$ est significative. L'hypothèse testée est donc :

- **H0 :** $\beta_1 = \beta_2 = ..... = \beta_p = 0$

contre

- **H1 :** il existe au moins un $\beta_j \neq 0$

On appelle *F* la statistique de test de Fisher-Snedecor, elle est donnée pour la régression multiple dans le tableau récapitulatif suivant :

| **Source de variation** | **Degré de liberté** | **Somme des carrés (des écarts)** | **Variance** | **F** |
|---|---|---|---|---|
| **Régression** | *p* | $SC_{régression}$ | $\frac{SC_{Régression}}{p}$ | $F = \frac{SC_{Régression}}{SC_{Résiduel}}$ |
| **Résiduel** | *n-p*-1 | $SC_{Résiduel}$ | $\frac{SC_{Résiduel}}{n-p-1}$ | |
| **Total** | *n*-1 | $SC_{Total}$ | | |

La table du test de Fisher-Snedecor fournit une valeur $F_{1-\alpha}(\nu_1, \nu_2)$ en fonction du nombre de degrés de liberté du numérateur $\nu_1$ (dans le cas de la régression multiple, $\nu_1 = p$), du nombre de degrés de liberté du dénominateur $\nu_2$ (dans le cas de la régression simple, $\nu_2 = n-p-1$) et de la valeur du risque $\alpha$.

Après avoir fixé la valeur du risque $\alpha$ , on conclut au rejet ou non-rejet de la significativité du modèle de la façon suivante :

1. Si la statistique *F* du test (calculée avec les données) est inférieure ou égale à $F_{1-\alpha}(\nu_1,\nu_2)$ l'hypothèse H0 est acceptée, c'est à dire que le modèle n'est pas significatif.

2. Si la statistique *F* du test est supérieure à $F_{1-\alpha}(\nu_1,\nu_2)$ l'hypothèse H0 est rejetée, c'est à dire que le modèle est significatif.

## TABLE DE FISHER-SNEDECOR

*Fractiles $F_{0,95}(\nu_1, \nu_2)$ de la loi de Fisher-Snedecor*

| DEGRÉS DE LIBERTÉ DU DÉNOMINATEUR : $\nu_2$ | DEGRÉS DE LIBERTÉ DU NUMÉRATEUR : $\nu_1$ | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 12 | 14 |
| 1 | 161 | 200 | 216 | 225 | 230 | 234 | 237 | 239 | 241 | 242 | 244 | 245 |
| 2 | 18,5 | 19,0 | 19,2 | 19,2 | 19,3 | 19,3 | 19,4 | 19,4 | 19,4 | 19,4 | 19,4 | 19,4 |
| 3 | 10,1 | 9,55 | 9,28 | 9,12 | 9,01 | 8,94 | 8,89 | 8,85 | 8,81 | 8,79 | 8,74 | 8,71 |
| 4 | 7,71 | 6,94 | 6,59 | 6,39 | 6,26 | 6,16 | 6,09 | 6,04 | 6,00 | 5,96 | 5,91 | 5,87 |
| 5 | 6,61 | 5,79 | 5,41 | 5,19 | 5,05 | 4,95 | 4,88 | 4,82 | 4,77 | 4,74 | 4,68 | 4,64 |
| 6 | 5,99 | 5,14 | 4,76 | 4,53 | 4,39 | 4,28 | 4,21 | 4,15 | 4,10 | 4,06 | 4,00 | 3,96 |
| 7 | 5,59 | 4,74 | 4,35 | 4,12 | 3,97 | 3,87 | 3,79 | 3,73 | 3,68 | 3,64 | 3,57 | 3,53 |
| 8 | 5,32 | 4,46 | 4,07 | 3,84 | 3,69 | 3,58 | 3,50 | 3,44 | 3,39 | 3,35 | 3,28 | 3,24 |
| 9 | 5,12 | 4,26 | 3,86 | 3,63 | 3,48 | 3,37 | 3,29 | 3,23 | 3,18 | 3,14 | 3,07 | 3,03 |
| 10 | 4,96 | 4,10 | 3,71 | 3,48 | 3,33 | 3,22 | 3,14 | 3,07 | 3,02 | 2,98 | 2,91 | 2,86 |
| 11 | 4,84 | 3,98 | 3,59 | 3,36 | 3,20 | 3,09 | 3,01 | 2,95 | 2,90 | 2,85 | 2,79 | 2,74 |
| 12 | 4,75 | 3,89 | 3,49 | 3,26 | 3,11 | 3,00 | 2,91 | 2,85 | 2,80 | 2,75 | 2,69 | 2,64 |
| 13 | 4,67 | 3,81 | 3,41 | 3,18 | 3,03 | 2,92 | 2,83 | 2,77 | 2,71 | 2,67 | 2,60 | 2,55 |
| 14 | 4,60 | 3,74 | 3,34 | 3,11 | 2,96 | 2,85 | 2,76 | 2,70 | 2,65 | 2,60 | 2,53 | 2,48 |
| 15 | 4,54 | 3,68 | 3,29 | 3,06 | 2,90 | 2,79 | 2,71 | 2,64 | 2,59 | 2,54 | 2,48 | 2,42 |
| 16 | 4,49 | 3,63 | 3,24 | 3,01 | 2,85 | 2,74 | 2,66 | 2,59 | 2,54 | 2,49 | 2,42 | 2,37 |
| 17 | 4,45 | 3,59 | 3,20 | 2,96 | 2,81 | 2,70 | 2,61 | 2,55 | 2,49 | 2,45 | 2,38 | 2,33 |
| 18 | 4,41 | 3,55 | 3,16 | 2,93 | 2,77 | 2,66 | 2,58 | 2,51 | 2,46 | 2,41 | 2,34 | 2,29 |
| 19 | 4,38 | 3,52 | 3,13 | 2,90 | 2,74 | 2,63 | 2,54 | 2,48 | 2,42 | 2,38 | 2,31 | 2,26 |
| 20 | 4,35 | 3,49 | 3,10 | 2,87 | 2,71 | 2,60 | 2,51 | 2,45 | 2,39 | 2,35 | 2,28 | 2,22 |
| 21 | 4,32 | 3,47 | 3,07 | 2,84 | 2,68 | 2,57 | 2,49 | 2,42 | 2,37 | 2,32 | 2,25 | 2,20 |
| 22 | 4,30 | 3,44 | 3,05 | 2,82 | 2,66 | 2,55 | 2,46 | 2,40 | 2,34 | 2,30 | 2,23 | 2,17 |
| 23 | 4,28 | 3,42 | 3,03 | 2,80 | 2,64 | 2,53 | 2,44 | 2,37 | 2,32 | 2,27 | 2,20 | 2,15 |
| 24 | 4,26 | 3,40 | 3,01 | 2,78 | 2,62 | 2,51 | 2,42 | 2,36 | 2,30 | 2,25 | 2,18 | 2,13 |
| 25 | 4,24 | 3,39 | 2,99 | 2,76 | 2,60 | 2,49 | 2,40 | 2,34 | 2,28 | 2,24 | 2,16 | 2,11 |
| 26 | 4,23 | 3,37 | 2,98 | 2,74 | 2,59 | 2,47 | 2,39 | 2,32 | 2,27 | 2,22 | 2,15 | 2,09 |
| 27 | 4,21 | 3,35 | 2,96 | 2,73 | 2,57 | 2,46 | 2,37 | 2,31 | 2,25 | 2,20 | 2,13 | 2,08 |
| 28 | 4,20 | 3,34 | 2,95 | 2,71 | 2,56 | 2,45 | 2,36 | 2,29 | 2,24 | 2,19 | 2,12 | 2,06 |
| 29 | 4,18 | 3,33 | 2,93 | 2,70 | 2,55 | 2,43 | 2,35 | 2,28 | 2,22 | 2,18 | 2,10 | 2,05 |
| 30 | 4,17 | 3,32 | 2,92 | 2,69 | 2,53 | 2,42 | 2,33 | 2,27 | 2,21 | 2,16 | 2,09 | 2,04 |
| 32 | 4,15 | 3,29 | 2,90 | 2,67 | 2,51 | 2,40 | 2,31 | 2,24 | 2,19 | 2,14 | 2,07 | 2,01 |
| 34 | 4,13 | 3,28 | 2,88 | 2,65 | 2,49 | 2,38 | 2,29 | 2,23 | 2,17 | 2,12 | 2,05 | 1,99 |
| 36 | 4,11 | 3,26 | 2,87 | 2,63 | 2,48 | 2,36 | 2,28 | 2,21 | 2,15 | 2,11 | 2,03 | 1,98 |
| 38 | 4,10 | 3,24 | 2,85 | 2,62 | 2,46 | 2,35 | 2,26 | 2,19 | 2,14 | 2,09 | 2,02 | 1,96 |
| 40 | 4,08 | 3,23 | 2,84 | 2,61 | 2,45 | 2,34 | 2,25 | 2,18 | 2,12 | 2,08 | 2,00 | 1,95 |
| 50 | 4,03 | 3,18 | 2,79 | 2,56 | 2,40 | 2,29 | 2,20 | 2,13 | 2,07 | 2,03 | 1,95 | 1,89 |
| 60 | 4,00 | 3,15 | 2,76 | 2,53 | 2,37 | 2,25 | 2,17 | 2,10 | 2,04 | 1,99 | 1,92 | 1,86 |
| 70 | 3,98 | 3,13 | 2,74 | 2,50 | 2,35 | 2,23 | 2,14 | 2,07 | 2,02 | 1,97 | 1,89 | 1,84 |
| 80 | 3,96 | 3,11 | 2,72 | 2,49 | 2,33 | 2,21 | 2,13 | 2,06 | 2,00 | 1,95 | 1,88 | 1,82 |
| 90 | 3,95 | 3,10 | 2,71 | 2,47 | 2,32 | 2,20 | 2,11 | 2,04 | 1,99 | 1,94 | 1,86 | 1,80 |
| 100 | 3,94 | 3,09 | 2,70 | 2,46 | 2,31 | 2,19 | 2,10 | 2,03 | 1,97 | 1,93 | 1,85 | 1,79 |
| ∞ | 3,84 | 3,00 | 2,60 | 2,37 | 2,21 | 2,10 | 2,01 | 1,94 | 1,88 | 1,83 | 1,75 | 1,69 |

Si on rejette H0, c'est qu'un ou plusieurs des coefficients sont significatifs. Il faut donc chercher quels sont ces coefficients à l'aide du test de Student sur chacun des coefficients.

### ***Le test de STUDENT***

On effectue aussi en régression linéaire multiple le test de Student qui va tester si chacun des coefficients estimés de la régression est significativement différent de zéro. La manière de procéder est la même qu'en régression simple, il faut donc se reporter à la fiche "QUALITE D'AJUSTEMENT" du modèle linéaire simple.

### ***Le test de FISHER-SNEDECOR partiel***

Le test de Fisher-Snedecor teste la signification de la régression dans sa globalité (il teste la nullité de tous les coefficients en même temps). Il ne permet donc pas de préjuger la signification particulière des coefficients pris isolément. C'est ce que fait le test de Student qui teste un à un la signification des coefficients.

Il arrive parfois qu'on veuille évaluer l'apport d'un groupe de variables. Par exemple, si on a le modèle de régression multiple suivant :

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_1^2 + \beta_3 X_2 + \beta_4 X_2^2 + \beta_5 X_2^3$$

et que l'on souhaite tester si globalement le degré 2 est significatif dans la régression, ni le test de Fisher, ni celui de Student ne pourront être utilisés. On utilisera alors le test de Fisher partiel (F-partiel). Il teste l'hypothèse :

**H0** : $\beta_2 = \beta_4 = 0$ contre **H1** : $\beta_2 \neq 0$ ou $\beta_4 \neq 0$

Nous n'allons pas détailler ce test mais vous pouvez avoir plus de renseignements au service statistique.

# OUTILS INFORMATIQUES

# LES OUTILS D'EXCEL POUR LA REGRESSION

## 1-DROITEREG (fonction d'Excel)

### DANS LE MENU "INSERTION" Sélectionner ⇒ Fonction : DROITEREG

Fonction EXCEL : **DROITEREG**

Calcule les statistiques pour une régression par la méthode des moindres carrés.

Formulation : **Droitereg(Y;X;VRAI;VRAI)**

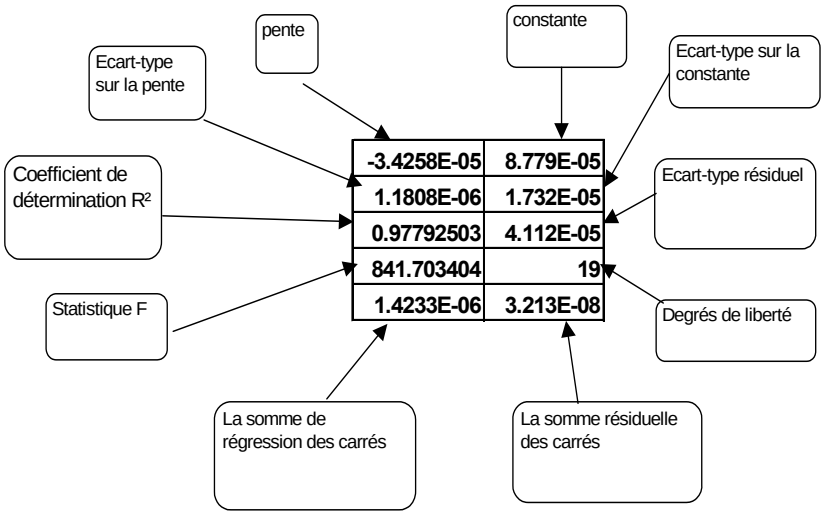
- pour afficher les statistiques associées.
- pour estimer la constante dans le modèle

**Procédure**: 1-Choisir la plage de sortie des résultats de la fonction (5lignes et 2 colonnes)

Fixer celle-ci en choisissant liste de choix (bouton droit de la souris)

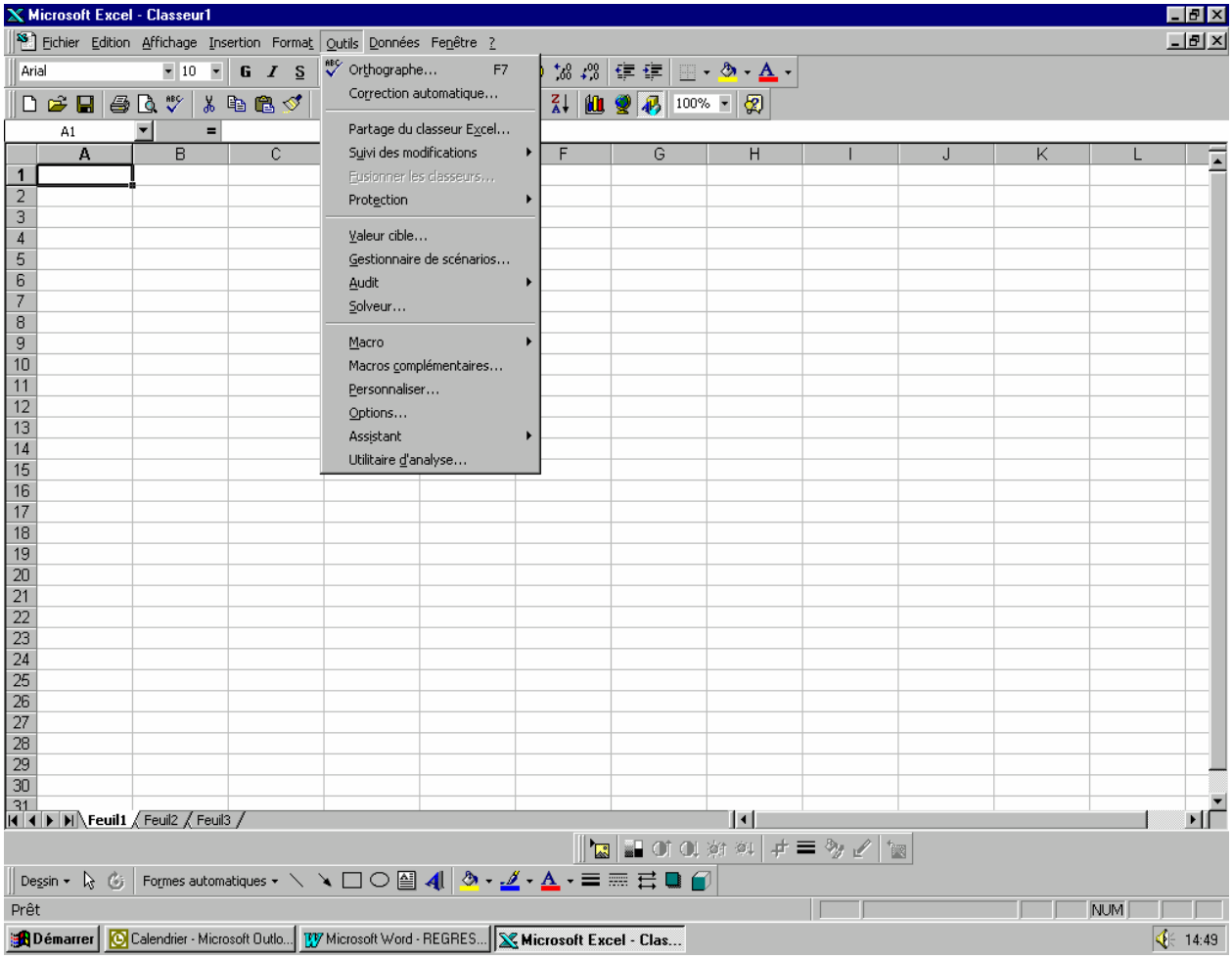
Taper la fonction droitreg (X;Y;Vrai;vrai) et validez par **CTRL+MAJ+ENTREE**

| N° | (mm) | (mm) |
|---|---|---|
| 1 | 0 | 0.000 100 000 |
| 2 | 1.255 | 0.000 055 000 |
| 3 | 2.51 | 0.000 005 000 |
| 4 | 3.765 | -0.000 025 000 |
| 5 | 5.02 | -0.000 110 000 |
| 6 | 6.275 | -0.000 090 000 |
| 7 | 7.53 | -0.000 140 000 |
| 8 | 8.785 | -0.000 145 000 |
| 9 | 10.04 | -0.000 295 000 |
| 10 | 11.295 | -0.000 300 000 |
| 11 | 12.55 | -0.000 400 000 |
| 12 | 13.805 | -0.000 400 000 |
| 13 | 15.06 | -0.000 500 000 |
| 14 | 16.315 | -0.000 490 000 |
| 15 | 17.57 | -0.000 570 000 |
| 16 | 18.825 | -0.000 580 000 |
| 17 | 20.08 | -0.000 605 000 |
| 18 | 21.335 | -0.000 605 000 |
| 19 | 22.59 | -0.000 710 000 |
| 20 | 23.845 | -0.000 655 000 |
| 21 | 25.1 | -0.000 725 000 |




Voir **Annexe C** pour plus de détails

## 2- REGRESSION LINEAIRE (Utilitaire d'analyse d'Excel)

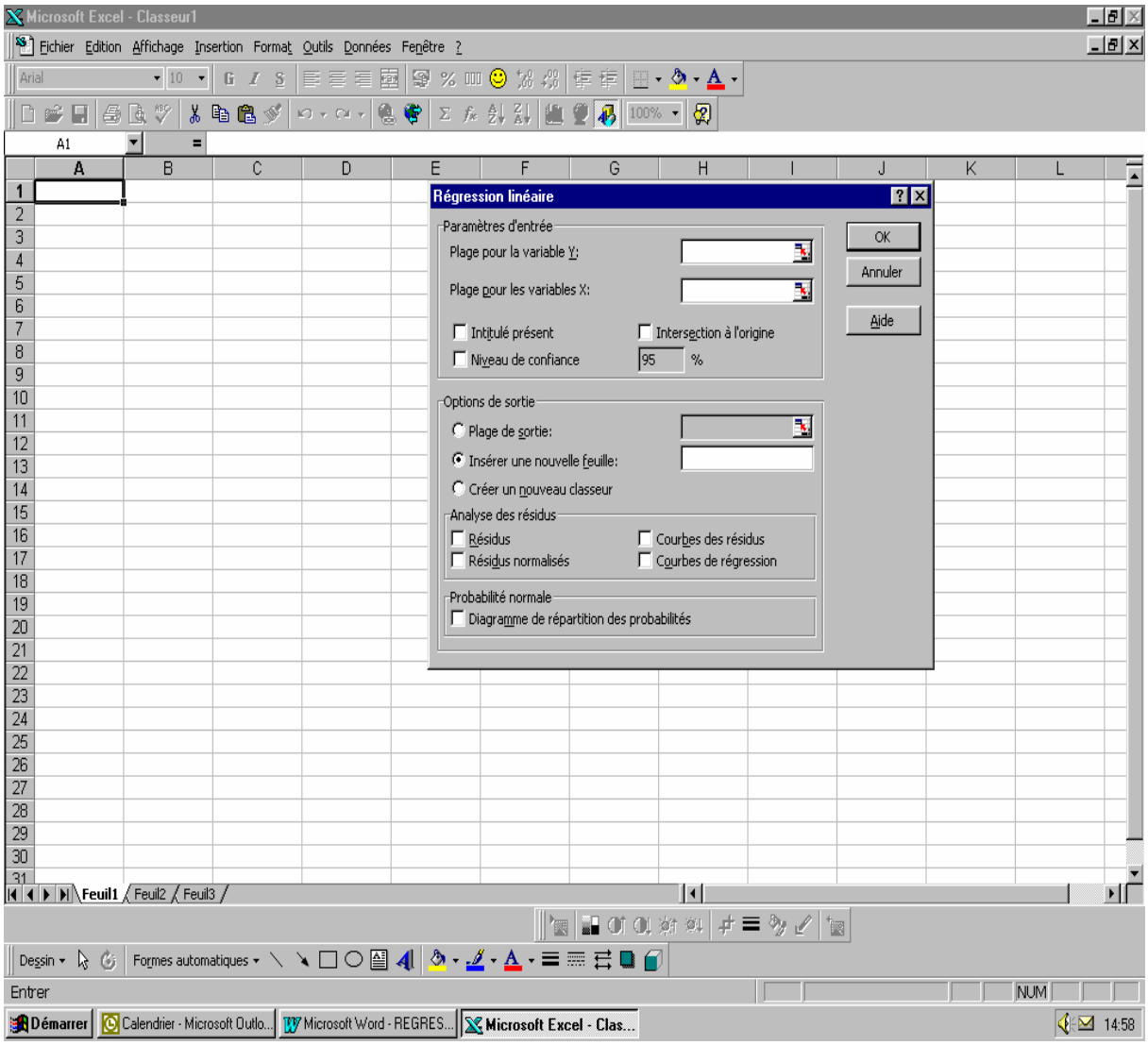



### 1 DANS LE MENU "OUTILS" CHOISIR ⇒"UTILITAIRE D'ANALYSE "

**Procédure :**

1- Si cette rubrique n'existe pas, l'installation s'effectue à partir de la rubrique "MACROS COMPLEMENTAIRE" du menu "OUTILS". Cochez les cas □ Utilitaire d'analyse et □ utilitaire d'analyse-VBA. Lancez l'installation en fermant par OK. La rubrique "UTILITAIRE" apparaît alors dans le menu "OUTILS".

2-Choisir REGRESSION LINEAIRE de la rubrique "Utilitaire d'analyse"




Suivre les indications du menu affiché :
Entrer la matrice des X et la matrice des Y
Cocher les informations utiles :
-Courbes de régression
-Résidus
**…**
**En validant par OK, vous obtiendrez les informations suivantes :** UTILITAIRE.XLS

| Statistiques de la régression | |
|---|---|
| Coefficient de détermination multiple (Coefficient de corrélation) (1) | 0.999947456 |
| Coefficient de détermination R^2 (2) & (3) | 0.999894913 |
| Coefficient de détermination R^2 (Coefficient ajusté) (4) | 0.999868641 |
| Erreur-type | 0.214419322 |
| Observations (nbre de réalisations) | 6 |

$$r = \sqrt{\frac{\sum (\hat{y}_i - \bar{y})^2}{\sum (y_i - \bar{y})^2}} \quad (1)$$

$$R^2 = \frac{\sum (\hat{y}_i - \bar{y})^2}{\sum (y_i - \bar{y})^2} \quad (2) \qquad R^2 = 1 - \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - \bar{y})^2} \quad (3)$$

$$R^2_{ajusté} = 1 - \frac{\sum (y_i - \hat{y}_i)^2 / N - 2}{\sum (y_i - \bar{y})^2 / N - p - 1} \quad (4)$$

**ANALYSE DE VARIANCE**

| | *Degré de liberté* | *Somme des carrés* | *Moyenne des carrés* | *F* | *Valeur critique de F* |
|---|---|---|---|---|---|
| Régression | 1 | 1749.8161 | 1749.8161 | 38059.6306 | 4.1414E-09 |
| Résidus | 4 | 0.18390258 | 0.04597565 | | |
| Total | 5 | 1750 | | | |

**TEST GLOBAL(test de Fisher) : Il permet de vérifier si il existe une relation linéaire entre X et Y**

**Ecart-type résiduel**

**Ecart-type de b0**

**Test de significativité de chaque coefficient dans la régression linéaire par le test de student. Si probabilité (P-Value) est supérieure à 0,05 --> le coefficient n'est pas significatif dans le modèle.**

| | *Coefficients* | *Erreur-type* | *Statistique t* | *Probabilité* | *Limite inférieure pour seuil de confiance = 95%* | *Limite supérieure pour seuil de confiance = 95%* | *Limite inférieure pour seuil de confiance = 95.0%* | *Limite supérieure pour seuil de confiance = 95.0%* |
|---|---|---|---|---|---|---|---|---|
| Constante | -0.15689887 | 0.15585563 | -1.00669364 | 0.37103683 | -0.58962435 | 0.27582662 | -0.58962435 | 0.27582662 |
| Variable X 1 | 0.000390266 | 2.0005E-06 | 195.088776 | 4.1414E-09 | 0.00038471 | 0.00039582 | 0.00038471 | 0.00039582 |

**Ecart-type de b1**

**Intervalle de confiance de b0 et b1 (affiché 2 fois)**

**ANALYSE DES RESIDUS**

**b0**

**b1**

| *Observation* | *Prévisions pour Y* | *Résidus* | *Résidus normalisés* |
|---|---|---|---|
| 1 | 0.164680666 | -0.16468067 | -0.85868466 |
| 2 | 10.01578572 | -0.01578572 | -0.08231056 |
| 3 | 19.79898442 | 0.20101558 | 1.04814364 |
| 4 | 29.97596194 | 0.02403806 | 0.12534023 |
| 5 | 39.78683711 | | 1.1316629 |
| 6 | 50.36415155 | | -1.36415155 |

**Ecart entre la valeur expériementale et la valeur modélisée**

**Voir la fiche sur les résidus**

**si F > valeur critique : il existe une relation statistiquement significative entre X et Y (la régression est significative)**

# LA REGRESSION PAR LA METHODE DE FORSYTHE :

**DOCUMENT 1 :** FORSYTHE.DOC

Ce document décrit le principe de la régression polynomiale selon la méthode de FORSYTHE.

**DOCUMENT 2 :** FORSYTHE_LNE.XLS

1- L'onglet « calculs » correspond à la feuille de base, avec les données à rentrer, les calculs, les graphes et les résultats.

2- L'onglet « résultats » donne juste un tableau avec les coefficients et leurs incertitudes, l'écart-type résiduel et la matrice de variance-covariance des coefficients.

3- L'onglet « résidus » contient une macro. Il permet de visualiser l'allure des résidus simple de la régression de *Y* sur les polynômes orthogonaux. (appuyer sur le bouton commande après avoir rentrer les données et le degré dans la feuille « calculs » pour avoir le graphe des résidus actualisé).

4- L'onglet « graphes » donne le graphe de la modélisation, ainsi que le graphe de l'incertitude due à la modélisation.

# Régression polynomiale - Polynômes orthogonaux - Méthode de Forsythe

## 1 MODELE SIMPLE

### 1.1 MODELE :

Le modèle théorique de départ est le suivant :

$Y = \sum_{j=0}^{k} \gamma_j X^j + \varepsilon$ n données $(X_i, Y_i)$

Ce n'est pas ce modèle sur lequel on travaille.

### 1.2 POLYNOMES ORTHOGONAUX :

Pour un tel modèle, l'utilisation de polynômes orthogonaux permet de minimiser les calculs et temps de calcul.

Lorsque les observations $X_i$ sont régulièrement espacées (ce qui est fréquent dans les séries chronologiques par exemple), il existe des tables de polynômes orthogonaux.

Lorsque le pas entre les observations $X_i$ n'est pas régulier, alors il faut construire les polynômes orthogonaux. La méthode de Forsythe est un moyen de générer une famille de tels polynômes (cf 1.3).

*Principe de l'utilisation des polynômes orthogonaux :*

Il s'agit de construire une famille de polynômes $P_j$ tels que :

- degré $(P_j)$=j
- $\forall$ (q,l) / q≠l, $\sum_{i=1}^{n} P_q(X_i) P_l(X_i) = 0$

Ensuite, on travaille sur le modèle $\boxed{Y = \sum_{i=0}^{k} \varphi_i P_i(X) + \eta}$ qui s'écrit, sous forme matricielle $Y = X\Phi + E$ avec $Var(E) = \sigma^2 I$

avec $X = \begin{bmatrix} P_0(X_1) & P_1(X_1) \ldots & P_k(X_1) \\ P_0(X_2) & P_1(X_2) \ldots & P_k(X_2) \\ & & \\ P_0(X_n) & P_1(X_n) \ldots & P_k(X_n) \end{bmatrix}$

$(X'X)^{-1}$ est alors une matrice diagonale de terme $A_{jj} = \dfrac{1}{\sum_{i=1}^{n} (P_j(X_i))^2}$

On obtient rapidement les estimations des coefficients de la régression de Y sur les $P_i(X)$ et de leur variance :

$$Y = \sum_{i=0}^{k} \hat{\varphi}_i P_i(X)$$

$$\hat{\varphi}_j = \frac{\sum_{i=1}^{n} Y_i P_j(X_i)}{\sum_{i=1}^{n} (P_j(X_i))^2} \qquad V(\hat{\varphi}_j) = \frac{\sigma^2}{\sum_{i=1}^{n} (P_j(X_i))^2} \qquad \mathrm{cov}(\hat{\varphi}_q, \hat{\varphi}_p) = 0 \quad si\ p \neq q$$

### 1.3 CONSTRUCTION DES POLYNOMES ORTHOGONAUX PAR LA METHODE DE FORSYTHE :

Dans la méthode de Forsythe, les polynômes vérifient le système d'équations suivant :

*Système {$S_X$}*

- $P_0(X) = 1$
- $P_1(X) = (X - \alpha_1)P_0(X)$
- $P_2(X) = (X - \alpha_2)P_1(X) - \beta_2 P_0(X)$
- ...
- $P_k(X) = (X - \alpha_k)P_{k-1}(X) - \beta_k P_{k-2}(X)$

avec

$$\alpha_0 = \beta_0 = \beta_1 = 0 \quad ; \qquad \alpha_j = \frac{\sum_{i=1}^{n} X_i \left(P_{j-1}(X_i)\right)^2}{\sum_{i=1}^{n} \left(P_{j-1}(X_i)\right)^2} \quad ; \qquad \beta_j = \frac{\sum_{i=1}^{n} \left(P_{j-1}(X_i)\right)^2}{\sum_{i=1}^{n} \left(P_{j-2}(X_i)\right)^2}$$

## 1.4 EXPRESSION DES POLYNOMES DE FORSYTHE COMME FONCTION DIRECTE DES PUISSANCES DE X

On a $Y = \sum_{i=0}^{k} \hat{\varphi}_i P_i(X)$.

Or on cherche les coefficients du modèle $Y = \sum_{j=0}^{k} \gamma_j X^j + \varepsilon$.

Il faut donc écrire les polynômes $P_i(X)$ en fonction linéaire des puissances de $X$ pour déterminer les estimateurs des coefficients $\gamma_j$.

Recherche les coefficients $\lambda_{i,j}$ tels que $P_i(X) = \sum_{j=0}^{i} \lambda_{i,j} X^j$ :

- $P_0(X) = 1 \Rightarrow \lambda_{0,0} = 1$
- $P_1(X) = (X - \alpha_1)P_0(X) \Rightarrow \quad \lambda_{1,0} = -\alpha_1 \quad \lambda_{1,1} = 1$
- $P_2(X) = (X - \alpha_2)P_1(X) - \beta_2 P_0(X) = X^2 - (\alpha_1 + \alpha_2)X + \alpha_1\alpha_2 - \beta_2 \Rightarrow \begin{cases} \lambda_{2,0} = \alpha_1\alpha_2 - \beta_2 \\ \lambda_{2,1} = \alpha_1 + \alpha_2 \\ \lambda_{2,2} = 1 \end{cases}$

On peut montrer par récurrence que les coefficients $\lambda_{i,j}$ vérifient le système d'équations suivant :

$$\begin{cases} \lambda_{0,0} = 1 \quad \lambda_{1,0} = -\alpha_1 \quad \lambda_{1,1} = 1 \\ \lambda_{i,0} = -\alpha_i \lambda_{i-1,0} - \beta_i \lambda_{i-2,0} \\ \lambda_{i,j} = \lambda_{i-1,j-1} - \alpha_i \lambda_{i-1,j} - \beta_i \lambda_{i-2,j} \quad 1 \le j \le i-2 \\ \lambda_{i,i-1} = \lambda_{i-1,i-2} - \alpha_i \\ \lambda_{i,i} = 1 \end{cases}$$

### 1.5 DETERMINATION DES $\hat{\gamma}_i = C_i$ ET DE LEUR MATRICE DE VARIANCE-COVARIANCE

Rappel : on cherche les coefficients $\hat{\gamma}_i = C_i$ du modèle $Y = \sum_{j=0}^{k} \gamma_j X^j + \varepsilon$

Cf 1.2 et 1.4 $Y = \sum_{i=0}^{k} \hat{\varphi}_i \left[ \sum_{j=0}^{i} \lambda_{i,j} X^j \right] = \sum_{i=0}^{k} \left[ \sum_{i=j}^{k} \hat{\varphi}_i \lambda_{i,j} \right] X^j$

$$\Rightarrow \; C_j = \sum_{i=j}^{k} \hat{\varphi}_i \lambda_{i,j}$$

$$\Rightarrow \; Var(C_j) = \sum_{i=j}^{k} \lambda_{i,j}{}^2 \, Var(\hat{\varphi}_i) \quad \text{et} \quad Cov(C_j, C_q) = \sum_{i=q}^{k} \lambda_{i,j} \lambda_{i,q} Var(\hat{\varphi}_i) \qquad pour \qquad j \prec q \le k$$

## 2 POLYNOMES ORTHOGONAUX ET METHODE DE FORSYTHE AVEC TRANSFORMATION DES DONNEES INITIALES

*Transformation des données*

La transformation suivante permet de travailler sur des données $U_k$ dont les valeurs sont comprises entre -2 et 2, ce qui améliore la précision des calculs numériques.

$$U_k = -2 + 4\frac{X_k - m}{M - m} \qquad \begin{cases} M = Max\{X_k\} \\ m = \min\{X_k\} \\ k \in [1, n] \end{cases}$$

Notons $H = \frac{M - m}{4}$ et $L = \frac{M + m}{4}$ .

### 2.1 MODELE :

$Y = \sum_{j=0}^{k} \gamma_j X^j + \varepsilon$ n données$\left(X_i, Y_i\right)$

## 2.2 POLYNOMES ORTHOGONAUX

On travaille sur le modèle $Y = \sum_{i=0}^{k} \phi_i Q_i(U) + \eta'$ qui s'écrit, sous forme matricielle $Y = Z\Psi + G$

avec $Z = \begin{bmatrix} Q_0(U_1) & Q_1(U_1). \ . \ . & Q_k(U_1) \\ Q_0(U_2) & Q_1(U_2). \ . \ . & Q_k(U_2) \\ & & \\ Q_0(U_n) & Q_1(U_n). \ . \ . & Q_k(U_n) \end{bmatrix}$

$(Z'Z)^{-1}$ est une matrice diagonale de terme $A_{jj} = \dfrac{1}{\sum_{i=1}^{n} (Q_j(U_i))^2}$

On obtient

$$Y = \sum_{i=0}^{k} \hat{\phi}_i Q_i(U)$$

$$\hat{\phi}_j = \frac{\sum_{i=1}^{n} Y_i Q_j(U_i)}{\sum_{i=1}^{n} (Q_j(U_i))^2} \qquad V(\hat{\phi}_j) = \frac{\sigma^2}{\sum_{i=1}^{n} (Q_j(U_i))^2} \qquad \mathrm{cov}(\hat{\phi}_q, \hat{\phi}_p) = 0 \quad \textit{si } p \neq q$$

## 2.3 CONSTRUCTION DES POLYNOMES ORTHOGONAUX PAR LA METHODE DE FORSYTHE :

*Système { $S_U$}*

- $Q_0(U) = 1$
- $Q_1(U) = (U - \alpha_1')Q_0(U)$
- $Q_2(U) = (U - \alpha_2')Q_1(U) - \beta_2' Q_0(U)$
- ...
- $Q_k(U) = (U - \alpha_k')Q_{k-1}(U) - \beta_k' Q_{k-2}(U)$

avec

$$\alpha_0' = \beta_0' = \beta_1' = 0 \quad ; \quad \alpha_j' = \frac{\sum_{i=1}^{n} U_i \left(Q_{j-1}(U_i)\right)^2}{\sum_{i=1}^{n} \left(Q_{j-1}(U_i)\right)^2} \quad ; \quad \beta_j' = \frac{\sum_{i=1}^{n} \left(Q_{j-1}(U_i)\right)^2}{\sum_{i=1}^{n} \left(Q_{j-2}(U_i)\right)^2}$$

## 2.4 EXPRESSION DES POLYNOMES DE FORSYTHE COMME FONCTION DE X

Les polynômes $P_i(X) = H^i Q_i(U)$ sont orthogonaux et générés par la méthode de Forsythe avec $\alpha_j = \alpha_j' H + 2L$, $\beta_j = \beta_j' H^2$.

## 2.5 EXPRESSION DES $P_i$ COMME FONCTION LINEAIRE DES PUISSANCES DE X

Après avoir défini les polynômes $P_i$ comme des polynômes générés par les équations de Forsythe, on peut reprendre les résultats obtenus en 1.4 , en tenant compte de la nouvelle définition des $\alpha_j$ et $\beta_j$.

$$\boxed{P_i(X) = \sum_{j=0}^{i} \mu_{i,j} X^j} \text{ avec } \begin{cases} \mu_{0,0} = 1 \quad \mu_{1,0} = -\alpha_1 \quad \mu_{1,1} = 1 \\ \mu_{i,0} = -\alpha_i \mu_{i-1,0} - \beta_i \mu_{i-2,0} \\ \mu_{i,j} = \mu_{i-1,j-1} - \alpha_i \mu_{i-1,j} - \beta_i \mu_{i-2,j} \qquad 1 \le j \le i-2 \\ \mu_{i,i-1} = \mu_{i-1,i-2} - \alpha_i \\ \mu_{i,i} = 1 \end{cases}$$

et $\alpha_j = \alpha_j' H + 2L, \quad \beta = \beta_j' H^2$

## 2.6 COMPARAISON AVEC LES RESULTATS DU PARAGRAPHE §I

rappel (cf §1) : $$Y = \sum_{i=0}^{k} \varphi_i P_i(X) + \eta \qquad (a)$$

$$\hat{Y} = \sum_{i=0}^{k} \hat{\varphi}_i P_i(X)$$

cf §2.2 et §4 : $$Y = \sum_{i=0}^{k} \phi_i Q_i(U) + \eta' \qquad (b)$$

$$\hat{Y} = \sum_{i=0}^{k} \hat{\phi}_i Q_i(U) = \sum_{i=0}^{k} \hat{\phi}_i \left( H^{-i} P_i(X) \right)$$

On a les égalités suivantes : $$\begin{cases} Q_i(U) = H^{-i} P_i(X) \\ \hat{\varphi}_i = H^{-i} \hat{\phi}_i \\ \eta = \eta' \\ \lambda_{i,j} = \mu_{i,j} \end{cases}$$

En fait, il suffit d'introduire le produit $H^i H^{-i}$ dans le modèle simple (a), pour obtenir le modèle (b) :

$$\hat{Y} = \sum_{i=0}^{k} \hat{\varphi}_i P_i(X) = \sum_{i=0}^{k} \hat{\varphi}_i \left( H^i H^{-i} \right) P_i(X) = \sum_{i=0}^{k} \left( \hat{\varphi}_i H^i \right) \left( H^{-i} P_i(X) \right) = \sum_{i=0}^{k} \hat{\phi}_i Q_i(U)$$

## 2.7 DETERMINATION DES $\hat{\gamma}_i = C_i$ ET DE LEUR MATRICE DE VARIANCE-COVARIANCE

On remplace $\hat{\varphi}_i$ par $H^{-i}\hat{\phi}_i$ et $\lambda_{i,j}$ par $\mu_{i,j}$ dans l'expression des coefficients $C_i$ dans 1.5

$$\Rightarrow \quad C_j = \sum_{i=j}^{k} \phi_i H^{-i} \ \mu_{i,j}$$

$$\Rightarrow Var(C_j) = \sum_{i=j}^{k} \frac{\mu_{i,j}^{\ 2}}{H^{2i}} Var(\hat{\phi}_i) \ \text{ et } \ Cov(C_j, C_q) = \sum_{i=q}^{k} \frac{\mu_{i,j}\,\mu_{i,q}}{H^{2i}} Var(\hat{\phi}_i) \quad pour \quad j \prec q \le k$$

# 3 POLYNOMES ORTHOGONAUX ET METHODE DE FORSYTHE AVEC TRANSFORMATION DES DONNEES INITIALES ET PONDERATION

## 3.1 REGRESSION PONDEREE - HETEROSCEDASTICITE - RAPPELS :

Ecriture matricielle $Y = X\beta + \varepsilon$
n données $\left(X_i, Y_i\right)$

*Hyp* : on a des informations sur les variances des $Y_i$ (ou des $\varepsilon_i$).

Par exemple, on connait V telle que matrice de variance-covariance de $\varepsilon$ = $V\sigma^2$.

$V = P'P = P^2$

$V^{-1} = (P^{-1})^2 = W = \left\{diag(w_i); i \in [1,n]\right\}$ matrice des poids

Nous pouvons appliquer la théorie des mco sur le modèle $P^{-1}Y = P^{-1}X\beta + P^{-1}\varepsilon$

L'estimation des moindres carrés de $\beta$ est donnée par $b = (X'V^{-1}X)^{-1}X'V^{-1}Y$.
L'estimation de la variance de $\beta$ est $Var(b) = (X'V^{-1}X)^{-1}\sigma^2$

Enfin, les résidus du modèle $P^{-1}Y = P^{-1}X\beta + P^{-1}\varepsilon$ sont $f = P^{-1}\varepsilon = P^{-1}(Y - \hat{Y})$
Leur variance est $Var(P^{-1}\varepsilon) = \sigma^2 I$

## 3.2 POLYNOMES ORTHOGONAUX

Le fait de pondérer les observations va entraîner des modifications dans les estimateurs des coefficients de régression de Y sur les polynômes orthogonaux, ainsi que dans la définition des polynômes orthogonaux.

On travaille sur le modèle $Y = \sum_{i=0}^{k} \phi_i Q_i(U) + \varepsilon'$ qui s'écrit, sous forme matricielle $Y = Z\Psi + G$
(cf. §2.2)

Si W est la matrice des poids définie en (3.1)

On obtient

$$\hat{\phi}_j = \frac{\sum_{i=1}^{n} w_i Y_i Q_j(U_i)}{\sum_{i=1}^{n} w_i (Q_j(U_i))^2} \qquad V(\hat{\phi}_j) = \frac{\sigma^2}{\sum_{i=1}^{n} w_i (Q_j(U_i))^2} \qquad \mathrm{cov}(\hat{\phi}_q, \hat{\phi}_p) = 0 \quad si\ p \neq q$$

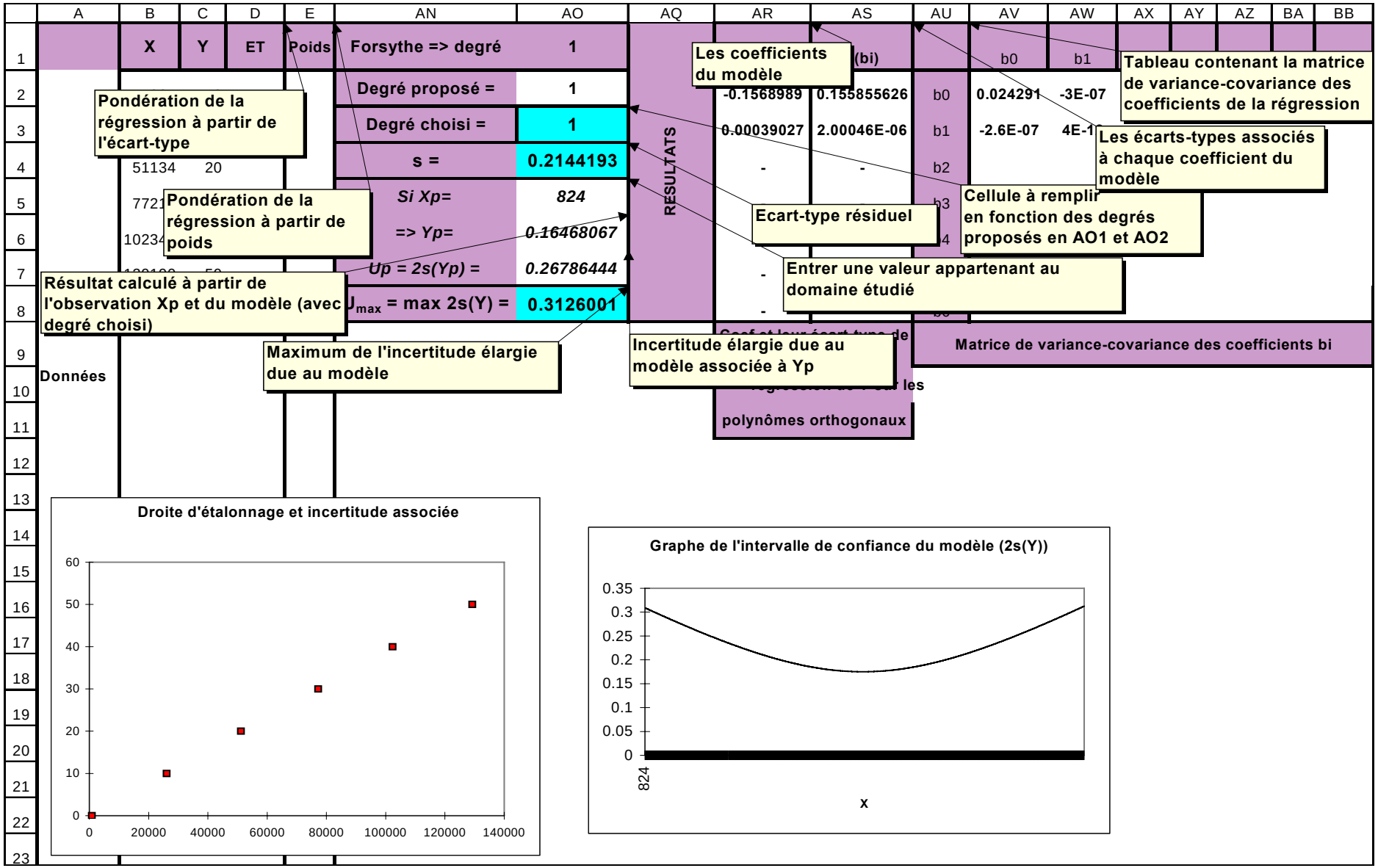
### 3.3 CONSTRUCTION DES POLYNOMES ORTHOGONAUX PAR LA METHODE DE FORSYTHE :

On reprend le système $\{S_U\}$ (cf 2.3) avec

$$\alpha_0' = \beta_0' = \beta_1' = 0 \quad ; \qquad \alpha_j' = \frac{\sum_{i=1}^{n} w_i U_i \left(Q_{j-1}(U_i)\right)^2}{\sum_{i=1}^{n} w_i \left(Q_{j-1}(U_i)\right)^2} \quad ; \qquad \beta_j' = \frac{\sum_{i=1}^{n} w_i \left(Q_{j-1}(U_i)\right)^2}{\sum_{i=1}^{n} w_i \left(Q_{j-2}(U_i)\right)^2}$$

Les étapes 4, 5 et 6 sont les mêmes que dans le paragraphe 2, à la différence de définition des $\alpha_j'$ et $\beta_j'$ près.

$$\Rightarrow \qquad C_j = \sum_{i=j}^{k} \phi_i H^{-i}\ \mu_{i,j}$$


X
Y
ET
Poids
Forsythe => degré
1
Degré proposé =
1
Degré choisi =
1
s =
0.2144193
Si Xp=
824
=> Yp=
0.16468067
Up = 2s(Yp) =
0.26786444
Umax = max 2s(Y) =
0.3126001
RESULTATS
Les coefficients du modèle
(bi)
-0.1568989
0.155855626
0.00039027
2.00046E-06
b0
b1
b2
b3
b0
b1
0.024291
-3E-07
-2.6E-07
4E-1
Tableau contenant la matrice de variance-covariance des coefficients de la régression
Les écarts-types associés à chaque coefficient du modèle
Pondération de la régression à partir de l'écart-type
Pondération de la régression à partir de poids
Cellule à remplir en fonction des degrés proposés en AO1 et AO2
Ecart-type résiduel
Entrer une valeur appartenant au domaine étudié
Résultat calculé à partir de l'observation Xp et du modèle (avec degré choisi)
Maximum de l'incertitude élargie due au modèle
Incertitude élargie due au modèle associée à Yp
Matrice de variance-covariance des coefficients bi
polynômes orthogonaux
Données
Droite d'étalonnage et incertitude associée
Graphe de l'intervalle de confiance du modèle (2s(Y))
824
x

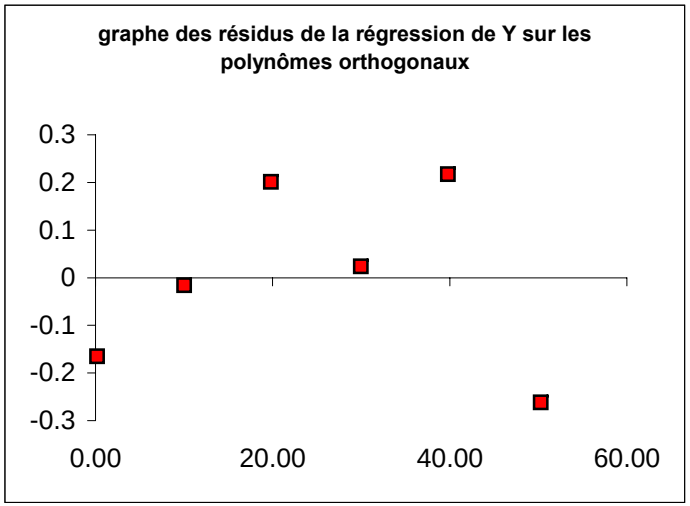
**LA REGRESSION POLYNOMIALE PAR LA METHODE DE FORSYTHE**

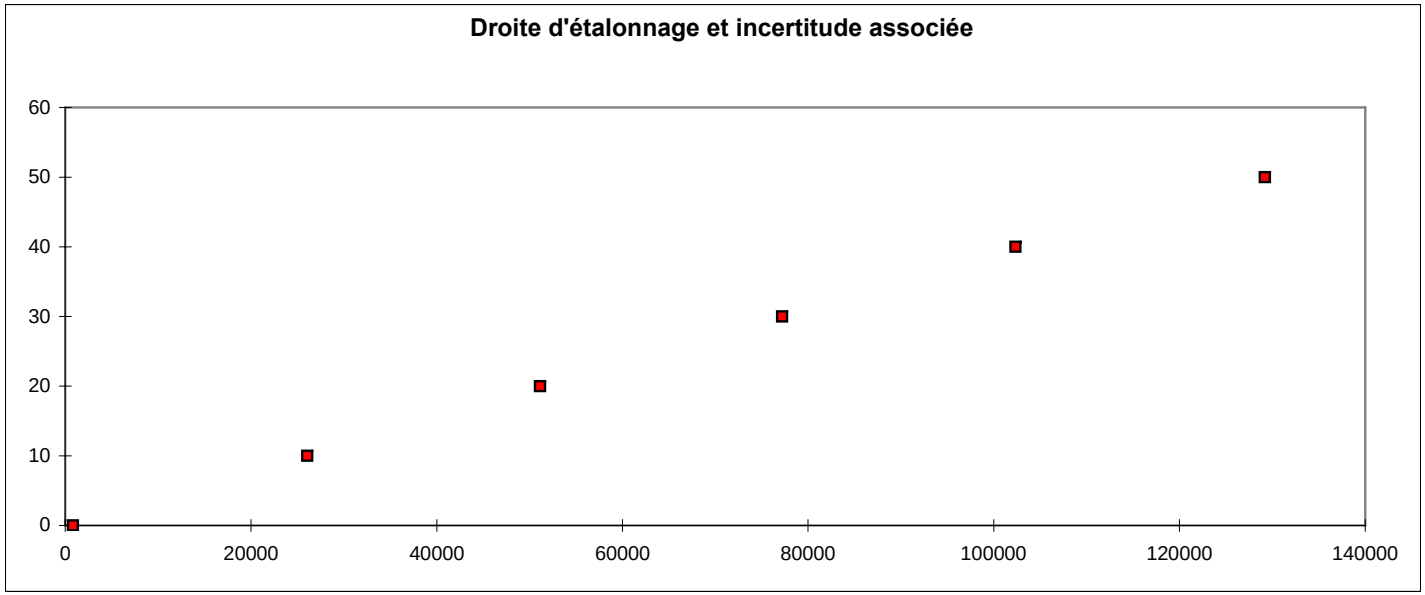
| RESULTATS | bi | s(bi) | s =Ecart-type résiduel | | b0 | b1 | b2 | b3 | b4 | b5 | b6 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | -0.15689887 | 0.15585563 | | b0 | 0.024290976 | -2.58E-07 | | | | | |
| | 0.00039027 | 2.0005E-06 | | b1 | -2.57961E-07 | 4.0018E-12 | | | | | |
| | - | - | | b2 | | | | | | | |
| | - | - | 0.214419322 | b3 | | | | | | | |
| | - | - | | b4 | | | | | | | |
| | - | - | | b5 | | | | | | | |
| | - | - | | b6 | | | | | | | |
| | Coef et leur écart-type de la régression de y sur les polynômes orthogonaux | | | | Matrice de variance-covariance des coefficients bi | | | | | | |

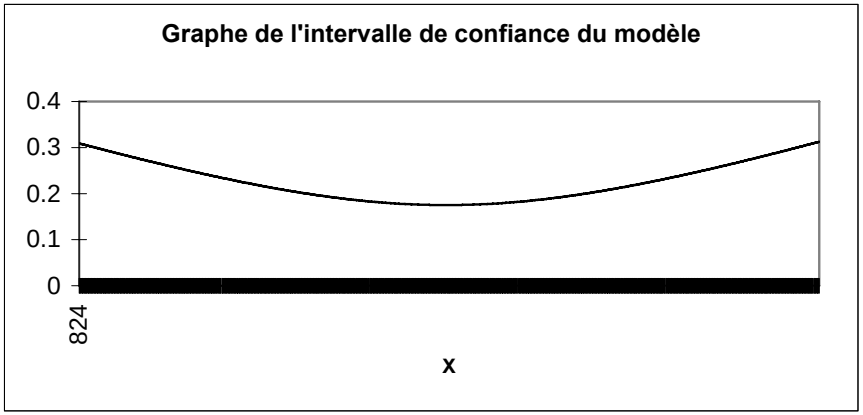
| Yajusté | résidus |
|---|---|
| 0.16 | -0.16468067 |
| 10.0157857 | -0.01578572 |
| 19.7989844 | 0.20101558 |
| 29.9759619 | 0.02403806 |
| 39.7829669 | 0.21703311 |
| 50.2616204 | -0.26162036 |

**Estimation des résidus simples**

6 GRAPHIQUE DES RESIDUS




Droite d'étalonnage et incertitude associée
60
50
40
30
20
10
0
0
20000
40000
60000
80000
100000
120000
140000


Graphe de l'intervalle de confiance du modèle
0.4
0.3
0.2
0.1
0
824
x

# CONCLUSION

Pour conclure, la régression est un outil très utilisé au LNE en particulier pour déterminer les courbes d'étalonnages, elle permet non seulement d'ajuster une courbe aux données, mais aussi de prédire de nouvelles valeurs et les incertitudes associées. Cependant, elle est à utiliser avec précaution, car elle repose sur des hypothèses en partie probabilistes qui, lorsqu'elles ne sont pas respectées conduisent à des résultats faux. Elle a de plus des limites, informatiquement les temps de calcul peuvent être longs dans le cas par exemple d'une régression polynomiale de degré élevé.

# ANNEXES

# REGRESSION LINEAIRE

## ANNEXE A : LES LEVIERS

*Définition des leviers* ( $h_{ii}$ )

Le calcul des résidus standardisés, press, studentisés (voir tableau « RESIDUS ») ainsi que les critères d'influence nécessitent l'introduction de nouvelles notations dont les leviers ( $h_{ii}$ ).

- Soient

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ . \\ . \\ . \\ Y_n \end{pmatrix} \qquad \mathbf{X} = \begin{pmatrix} X_1 \\ X_2 \\ . \\ . \\ . \\ X_n \end{pmatrix}$$

un échantillon de taille *n* des variables *Y* et *X* .
Sous forme matricielle, le modèle de régression linéaire s'écrit :

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

avec :

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ . \\ . \\ . \\ Y_n \end{pmatrix}, \ \mathbf{X} = \begin{pmatrix} 1 & X_1 \\ 1 & X_2 \\ . & . \\ . & . \\ . & . \\ 1 & X_n \end{pmatrix}, \ \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}, \ \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ . \\ . \\ . \\ \varepsilon_n \end{pmatrix}$$

On peut alors écrire l'estimateur de $\beta$ sous la forme $\mathbf{B} = \mathbf{(X'X)^{-1}(X'Y)}$ où $\mathbf{B} = \begin{pmatrix} b_0 \\ b_1 \end{pmatrix}$

Et on a alors l'estimation suivante pour *Y* :

$$\begin{aligned} \hat{\mathbf{Y}} &= \mathbf{XB} \\ &= \mathbf{(X)(X'X)^{-1}(X'Y)} \\ &= \mathbf{HY} \end{aligned}$$

Où X' est la transposée de **X** et **H** est la «matrice chapeau » ( « hat matrice »).

On écrit **H** de la manière suivante :

$$\mathbf{H} = \mathbf{X(X'X)^{-1}X'} = \begin{pmatrix} h_{11}\, h_{12} & \ldots\ldots\ldots & h_{1n} \\ h_{21}\, h_{22} & \ldots\ldots\ldots & h_{2n} \\ \ldots\ldots\ldots & & \\ \ldots\ldots\ldots & & \\ h_{n1}\, h_{n2} & \ldots\ldots\ldots & h_{nn} \end{pmatrix}$$

Dans le cadre de la régression simple, on a : $h_{ij} = \frac{1}{n} + \frac{(X_i - \overline{X})(X_j - \overline{X})}{\sum_{i=1}^{n}(X_j - \overline{X})^2}$

donc en particulier pour *i=j* , on a : $h_{ii} = \frac{1}{n} + \frac{(X_i - \overline{X})^2}{\sum_{i=1}^{n}(X_j - \overline{X})^2} \quad \forall i \in \{1, n\}$

Les valeurs $h_{ii}$ sont appelées « leviers » ou en anglais « leverage ». Ce sont les éléments diagonaux de la « hat matrice » et ils ont un rôle important dans le calcul des résidus et dans la détection de valeurs influentes.

A noter :

- La moyenne des leviers vaut $\frac{2}{n}$
- $0 \le h_{ii} \le 1 \qquad \forall i \in \{1, n\}$

Remarques :

$\hat{\mathbf{Y}} = \mathbf{HY}$ donc on a : $\begin{pmatrix} \hat{Y}_0 \\ \hat{Y}_1 \\ \vdots \\ \hat{Y}_n \end{pmatrix} = \begin{pmatrix} h_{11} & h_{12} & \cdots & h_{n1} \\ h_{21} & \ddots & & \vdots \\ \vdots & & \ddots & \vdots \\ h_{n1} & \cdots & \cdots & h_{nn} \end{pmatrix} \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}$ c'est à dire en développant,

$\forall i \in \{1, n\} \quad \hat{Y}_i = h_{i1}Y_1 + h_{i2}Y_2 + \ldots\ldots + h_{ii}Y_i + \ldots\ldots + h_{in}Y_n \qquad \forall \mathrm{i} \in \{1, \mathrm{n}\}$

$h_{ii}$ est donc le coefficient qui exprime l'importance de l'observation $Y_i$ pour l'estimation $\hat{Y}_i$. Si $h_{ii}$ est grand cela signifie donc que la ième observation $Y_i$ est influente (une faible variation de $Y_i$ induit une modification importante de la régression.)

# REGRESSION LINEAIRE

## ANNEXE B

AFNl NF X 06-050 95 ■ 1012372 0600268 622 ■

ISSN 0335-3931

norme française

**NF X 06-050**
**Décembre 1995**

Indice de classement : **X 06-050**

**ICS : 03.120.30**

Application de la statistique

## Étude de la normalité d'une distribution

E : Application of statistics — Study of the normality of a distribution
D : Angewandte Statistik — Tests auf Normalverteilung

© AFNOR 1995 — Tous droits réservés

### *Norme française homologuée*

par décision du Directeur Général de l'AFNOR le 20 novembre 1995 pour prendre effet le 20 décembre 1995.

Remplace la norme homologuée NF X 06-050, de décembre 1991.

### *Correspondance*

À la date de publication du présent document, il existe un projet de norme internationale ISO/DIS 5479 traitant du même sujet.

### *Analyse*

Le présent document fournit un choix de tests permettant de vérifier la validité de l'hypothèse de normalité d'une distribution. Il précise les conditions d'utilisation et les règles d'interprétation propres à chaque test.

### *Descripteurs*

**Thésaurus International Technique** : statistique, distribution statistique, test statistique, calcul des probabilités, méthode graphique.

### *Modifications*

Par rapport à la précédente édition, les modifications suivantes ont été apportées :

— le document a été restructuré pour introduire les graphiques aux endroits appropriés dans le texte et reporter en annexe les tests directionnels et le test conjoint en $\sqrt{b_1}$ et $b_2$.

— l'exemple de Test de Kolmogorov-Smirnov a été modifié.

### *Corrections*

Éditée et diffusée par l'Association Française de Normalisation (AFNOR), Tour Europe 92049 Paris La Défense Cedex — Tél. : (1) 42 91 55 55

© AFNOR 1995 AFNOR 1995 1er tirage 95-12

AFNI NF X 06-050 95 ■ 1012372 0600269 569 ■

## Méthodes statistiques AFNOR **X06E**

**Membres de la commission de normalisation**

Président : M BRUNSCHWIG

Secrétariat : MME DEL CERRO — AFNOR

| | | |
|---|---|---|
| M | BALLAUD | QUALITE SYSTEME |
| M | BARBIER | AEROSPATIALE |
| MME | BEGUERE | SLP STATISTIQUES |
| MME | BOUVENOT | AFNOR |
| M | BRUNSCHWIG | |
| M | CAILLOUX | ECOLE SUPERIEURE DE METROLOGIE |
| M | CAZALBOU | FRANCE TELECOM |
| M | CHEROUTE | PREVOYANCE SYSTEMES |
| M | DAUDIN | INAPG |
| MME | DESENFANT | LNE |
| M | ETIENNE | DAEI/METT |
| M | FEINBERG | CNEVA/CIQUAL |
| M | JAMBU | FRANCE TELECOM |
| M | KOLUB | SGS QUALITEST |
| M | LEGEAY | LCPC |
| MME | OUDIN DARRIBERE | |
| M | PALSKY | RHONE POULENC CHIMIE |
| M | PERRUCHET | UTAC |
| M | SADO | TOTAL RAFFINAGE DISTRIBUTION |
| M | SAPORTA | CNAM |
| M | SCHNELLBACH | PMS LOGICIEL |
| M | SUARD | ALLIED SIGNAL SYSTEMS DE FREINAGE SA |
| M | TOUTAIN | SCHNEIDER ELECTRIC SA |
| M | TUFFERY | CNEVA |
| M | WENISCH | SQIFE |
| M | ZANKEVITCH | DRIRE |

## Sommaire

Page

**0** **Introduction** ..... 4

**1** **Domaine d'application** ..... 4

**2** **Généralités** ..... 5

**3** **Méthode graphique** ..... 6
**3.1** Principe de cette méthode ..... 6
**3.2** Utilisation des fréquences cumulées (graphique de Henry) ..... 6
**3.2.1** Mode opératoire pour de petits échantillons (par exemple, $n < 30$) ..... 7
**3.2.2** Mode opératoire pour de grands échantillons ..... 10
**3.3** Observations générales relatives aux méthodes graphiques ..... 10
**3.3.1** Estimations ..... 10
**3.3.2** Changements de variables ..... 10
**3.3.3** Mélange de distributions ..... 13
**3.3.4** Valeurs aberrantes ..... 15

**4** **Tests omnibus** ..... 15
**4.1** Généralités ..... 15
**4.2** Test de Shapiro-Wilk ..... 16
**4.2.1** Description ..... 16
**4.2.2** Mode opératoire ..... 16
**4.2.3** Exemple ..... 17
**4.3** Test de Kolmogorov-Smirnov ..... 18
**4.3.1** Description ..... 18
**4.3.2** Mode opératoire ..... 18

**Annexe A** (normative) **Tests directionnels et test conjoint en $\sqrt{b_1}$ et $b_2$** ..... 20
**A.1** Tests directionnels ..... 20
**A.1.1** Généralités ..... 20
**A.1.2** Test directionnel d'asymétrie en $\sqrt{b_1}$ : $(8 \leq n \leq 5\,000)$ ..... 21
**A.1.3** Test directionnel d'aplatissement en $b_2$ : $(7 \leq n \leq 1\,000)$ ..... 23
**A.2** Test conjoint en $\sqrt{b_1}$ et $b_2$ (multidirectionnel) : $(20 \leq n \leq 1\,000)$ ..... 24
**A.2.1** Généralités ..... 24
**A.2.2** Test conjoint en $\sqrt{b_1}$ et $b_2$ ..... 24

**Annexe B** (normative) **Tables** ..... 26

**Annexe C** (informative) **Bibliographie** ..... 35

**Annexe D** (informative) **Exemple de papier gausso-arithmétique** ..... 36

# REGRESSION LINEAIRE

## ANNEXE C : DROITEREG D'EXCEL

### DROITEREG

Utilise la méthode des « moindres carrés » pour calculer une droite ajustée au plus près à vos données et renvoie une matrice qui décrit cette droite. L'équation de la droite est :

$y = m_1x_1 + m_2x_2 + ... + b$ ou $y = mx + b$

où la variable dépendante $y$ est une fonction des variables indépendantes $x$. Les valeurs $m$ sont des coefficients correspondant à chaque valeur $x$, et $b$ est une valeur constante. A noter que $y$, $x$ et $m$ peuvent être des vecteurs. La matrice renvoyée par la fonction DROITEREG a la forme $\{m_n.m_{n-1}.....m_1.b\}$. DROITEREG peut également renvoyer d'autres grandeurs statistiques de régression.

#### Syntaxe

DROITEREG(matrice_y;*matrice_x;const;stats*)

**Matrice_y** est la série des valeurs $y$ déjà connues par la relation $y = mx + b$.

- Si la *matrice_y* occupe une seule colonne, chaque colonne de la *matrice_x* est interprétée comme étant une variable séparée.
- Si la *matrice_y* occupe une seule ligne, chaque ligne de *matrice_x* est interprétée comme une variable séparée.

**Matrice_x** est une série de valeurs $x$ facultatives, éventuellement déjà données par la relation $y = mx + b$.

- La *matrice_x* peut contenir une ou plusieurs séries de variables. Si vous utilisez une seule variable, les formes de la *matrice_y* et de la *matrice_x* peuvent être différentes, mais elles doivent avoir la même dimension. Si vous utilisez plusieurs variables, la *matrice_y* doit être un vecteur (en d'autres termes, une plage de cellules dont la hauteur est une seule ligne et la largeur, une seule colonne).
- Si la *matrice_x* est omise, la matrice par défaut sera la matrice {1.2.3....} de même dimension que la *matrice_y*.

**Const** est une valeur logique qui indique si la constante *b* doit être égale à 1.

- Si *const* est VRAI ou omis, *b* est calculé normalement.
- Si *const* est FAUX, *b* est égal à 1 et les valeurs *m* sont ajustées de sorte que $y = mx$.

**Stats** est une valeur logique qui indique si d'autres grandeurs statistiques de régression doivent être renvoyées.

- Si *stats* est VRAI, DROITEREG renvoie des grandeurs de régression supplémentaires, sous la forme d'une matrice :

  $\{m_n,m_{n-1},\ldots;m_1;b;se_n,se_{n-1},\ldots..se_1,se_b;r^2,se_y;F,df;ss_{rég},ss_{résid}\}$.
- Si *stats* est FAUX ou omis, DROITEREG renvoie uniquement les coefficients *m* et la constante *b*.

Les grandeurs statistiques de régression supplémentaires sont :

| *Statistique* | *Description* |
|---|---|
| $se_1,se_2,\ldots,se_n$ | Les erreurs-types pour les coefficients $m_1,m_2,\ldots,m_n$. |
| $se_b$ | L'erreur-type pour la constante *b* ($se_b$ = #NA lorsque *const* est FAUX). |
| $r^2$ | Le coefficient de détermination. Compare les valeurs *y* estimées et les valeurs *y* réelles et fournit le résultat sur une échelle de 0 à 1. Si le coefficient vaut 1, l'échantillon est parfaitement en corrélation et il n'y a pas de corrélation entre les estimations de *y* et ses valeurs réelles. Inversement, si le coefficient de détermination vaut 0, l'équation de régression n'est pas utile pour la prédiction d'une valeur *y*. Pour plus d'informations sur le mode de calcul de $r^2$, voir les remarques ci-dessous. |
| $se_y$ | L'erreur-type pour l'estimation de *y*. |
| *F* | La statistique *F*, ou la valeur *F*-observée. Utilisez ce paramètre pour déterminer si la relation observée entre variables dépendantes et indépendantes est due au hasard. |
| df | Les degrés de liberté. Utilisez les degrés de liberté pour vous aider à trouver les valeurs critiques de la statistique *F* dans une table statistique. Comparez les valeurs que vous trouvez dans la table à la statistique *F* renvoyée par DROITEREG afin de déterminer le degré de confiance du modèle. |
| $ss_{rég}$ | La somme de régression des carrés. |
| $ss_{résid}$ | La somme résiduelle des carrés. |

L'illustration suivante montre l'ordre dans lequel les régressions statistiques supplémentaires sont renvoyées :

| | | | | | |
|---|---|---|---|---|---|
| $mn$ | $m_{n-1}$ | . . . | $m2$ | $m1$ | $b$ |
| $se_n$ | $se_{n-1}$ | . . . | $se_2$ | $se_1$ | $se_b$ |
| $r^2$ | $se_y$ | | | | |
| $F$ | df | | | | |
| $ss_{rég}$ | $ss_{résid}$ | | | | |

### Remarques

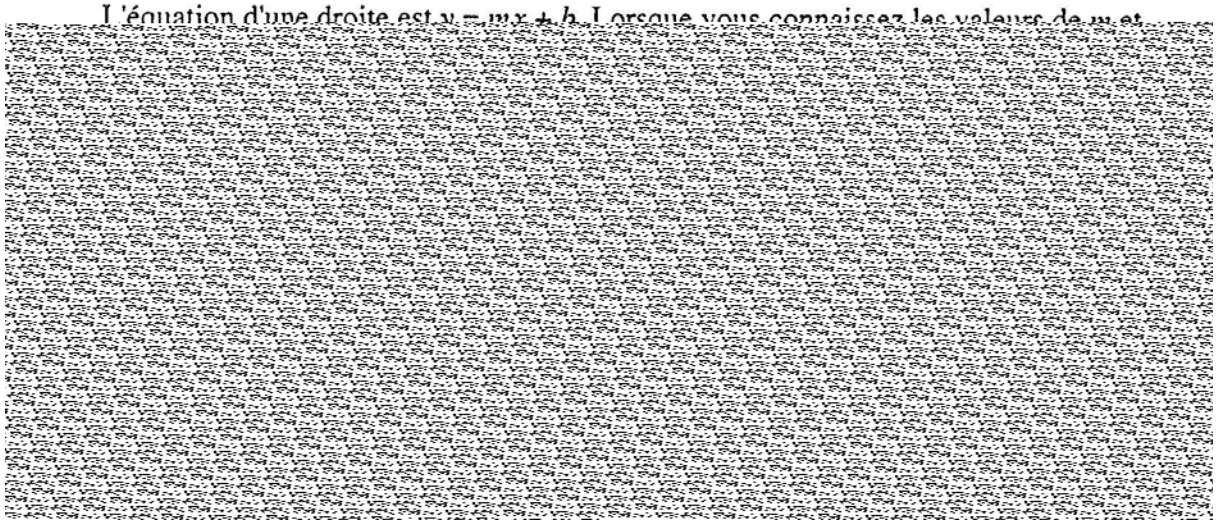
- Toute droite peut être décrite par sa pente et son intersection avec l'axe $y$ (ordonnées) :

  Pente ($m$) :

  Pour calculer la pente d'une droite, habituellement désignée par $m$, prenez deux points de la droite, $(x_1,y_1)$ et $(x_2,y_2)$ ; la pente est égale à : $(y_2 - y_1)/(x_2 - x_1)$.

  Intersection de la droite avec l'axe des ordonnées ($b$) :
  L'intersection d'une droite avec l'axe des ordonnées, habituellement désignée par $b$, est la valeur de $y$ au point d'intersection de la droite avec l'axe des ordonnées.

  L'équation d'une droite est $y = mx + b$. Lorsque vous connaissez les valeurs de $m$ et



« Entrée réservée » correspond à une entrée réservée aux fournisseurs. Lorsqu'elle est introduite sous forme de matrice, la formule suivante :

DROITEREG(E2:E12:A2:D12;VRAI;VRAI)

Donne le résultat suivant :

| | A | B | C | D | E |
|---|---|---|---|---|---|
| 14 | -2342,3716 | 25532,1066 | 125297,682 | 276,413874 | 523178,31 |
| 15 | 132,680115 | 5306,69152 | 4000,66838 | 54,2937404 | 122373,62 |
| 16 | 0,99674799 | 9705,78463 | #N/A | #N/A | #N/A |
| 17 | 459,753674 | 6 | #N/A | #N/A | #N/A |
| 18 | 1732393319 | 565213532 | #N/A | #N/A | #N/A |

L'équation de la régression multiple, $y = m_1{*}x_1 + m_2{*}x_2 + m_3{*}x_3 - m_4{*}x_4 + b$ peut, dès lors, être calculée au moyen des valeurs de la ligne 14 :

$y = 27{,}64{*}x_1 + 12\,530{*}x_2 + 2\,553{*}x_3 - 234{,}24{*}x_4 + 52\,318$

Le promoteur est maintenant en mesure d'estimer la valeur immobilière d'un immeuble de bureaux dans le même quartier, qui est vieux de 25 ans, occupe 2 500 mètres carrés de surface utile, trois bureaux et deux entrées, au moyen de l'équation suivante :

$y = 27{,}64{*}2500 + 12530{*}3 + 2553{*}2 - 234{,}24{*}25 + 52318 = 158\,260$ F

La fonction TENDANCE peut aussi être utilisée pour calculer cette valeur. Pour plus d'informations, voir la fonction TENDANCE, plus loin dans ce manuel.

### Exemple 4 : Statistiques *F* et $R^2$

Dans l'exemple précédent, le coefficient de détermination $r^2$ est 0,99675 (voir la cellule A16 dans le résultat de DROITEREG), ce qui laisse supposer une relation très étroite entre les variables indépendantes et le prix de vente. Vous pouvez utiliser la statistique *F* pour déterminer si la relation étroite des résultats qu'implique la valeur élevée de $r^2$ est le fruit du hasard.

Supposons un instant qu'il n'existe pas de relation véritable entre les variables, mais que l'échantillon de 11 immeubles de bureaux choisi est tel que l'analyse statistique montre une forte relation. La probabilité de conclure erronément qu'il existe bel et bien une relation est appelée Alpha.

Il existe une relation entre les variables si la statistique *F* observée est supérieure à une valeur critique donnée. Si vous devez connaître la valeur critique de la statistique *F*, consultez un manuel de statistiques ; la plupart fournissent une table des valeurs critiques de la statistique *F*. Pour lire la table, faites un test unilatéral, utilisez une valeur Alpha de 0,05, et pour les degrés de liberté (représentés par *v*1 et *v*2 dans la plupart des manuels), utilisez $v1 = k = 4$ et $v2 = n - (k + 1) = 11 - (4 + 1) = 6$, où *k* est le nombre de variables indépendantes utilisées dans la régression, et *n*, le nombre de points. La valeur critique de la statistique *F* est 4,53.

La valeur *F* observée est 459,753674 (cellule A17) et est nettement supérieure à la valeur critique de la statistique *F*, 4,53. Aussi, l'équation de régression est utile pour prévoir la valeur immobilière des immeubles de bureaux dans ce quartier.

**Exemple 5 : Calculer la statistique *t***

Un autre test d'hypothèse déterminera si chaque coefficient de pente est utile dans l'estimation de la valeur immobilière d'un immeuble de bureaux de l'exemple 3. Par exemple, en vue de tester le coefficient d'âge et de déterminer sa signification statistique, divisez –234.24 (coefficient de la pente âge) par 13.268 (l'erreur-type estimée des coefficients d'âge dans la cellule A15). Voici la valeur observée :

$t = m_4 + se_4 = -234.24 + 13.268 = -17.7$

Si vous consultez une table dans un manuel de statistiques, vous trouverez que la valeur critique *t* unilatérale, avec 6 degrés de liberté et Alpha = 0,05, est égale à 1,94. Etant donné que la valeur absolue de *t*, 17,7, est supérieure à 1,94, l'âge est une valeur significative pour l'estimation de la valeur immobilière d'un immeuble de bureaux. Chacune des autres variables indépendantes peut être testée de la même manière pour la détermination de sa signification statistique. Voici la liste des valeurs *t* observées pour chacune des variables indépendantes :

| | |
|---|---|
| Superficie utile | 5.1 |
| Nombre de bureaux | 31.3 |
| Nombre d'entrées | 4.8 |
| Age | –17.7 |

Toutes ces valeurs ont une valeur absolue supérieure à 1.94. Par conséquent, toutes les variables utilisées dans l'équation de régression sont utiles dans la prévision de la valeur immobilière des immeubles de bureaux de ce quartier.

**Fonctions corrélatives**

| *Fonction* | *Description* |
|---|---|
| CROISSANCE et LOGREG | Similaires à TENDANCE et DROITEREG, mais ajustent une courbe exponentielle et non une droite à vos données |
| DONNEES.SERIE | Recopie une série de cellules avec des valeurs interpolées, croissantes ou décroissantes |
| RECOPIE.INCREMENT | Recopie automatiquement une série de cellules en fonction de la taille et du contenu de la sélection |
| TENDANCE | Calcule une droite, mais renvoie une matrice des valeurs y prévues et non des paramètres de la droite |

Un exemple utilisant la fonction droitereg et montrant les différentes possibilités de graphes à associer, est traité sur le site suivant :

http://physinfo.ulb.ac.be/!UVA!/Informatique/!UVA!bureautique/Excel_5/Excel_5_33.html#Heading129

# REGRESSION LINEAIRE

## ANNEXE D : REGRESSION POLYNOMIALE

### LA REGRESSION POLYNOMIALE

Il semble important de faire une annexe sur la régression polynomiale, car c'est un type de modèle très souvent utilisé lorsque la régression simple n'est pas appropriée. Nous ne développons pas un chapitre à son sujet car elle n'est qu'un cas particulier de la régression linéaire multiple. En effet, pour un modèle à une variable explicative *X,* le modèle polynomial s'écrit :

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + .... + \beta_p X^p + \varepsilon$$

où p est le degré du polynôme. On dit alors que le modèle est d'ordre p.
Ce qu'il y a d'intéressant avec la régression polynomiale, c'est que l'on trouve souvent dans les logiciels de statistiques des algorithmes permettant de trouver le meilleur degré pour modéliser notre phénomène. Un critère de choix est par exemple, le modèle qui donnera le meilleur $R^2_{ajusté}$ .

La plupart du temps, les logiciels ne permettent pas d'utiliser des polynômes ayant un degré supérieur à 7 ou 8, car la minimisation de la somme des moindres carrés conduit à un système d'équations difficiles à résoudre. On peut alors pour remédier à ce problème utiliser une approximation polynomiale par les polynômes de forsythe. (VOIR FEUILLE EXCEL FORSYTHE)

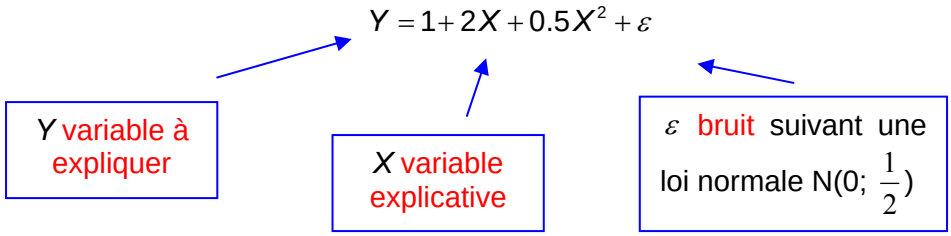
# REGRESSION LINEAIRE

## ANNEXE E : EXEMPLE DE REGRESSION

### *EXEMPLE : impact d'une valeur atypique*

#### Construction des données

Pour mettre en évidence l'influence d'une valeur atypique sur un ajustement par régression linéaire, nous avons construit un échantillon. Nous l'avons choisi de taille 50, avec une unique variable explicative *X* prenant pour valeurs les nombres entiers de 1 à 50 sans répétition. Nous avons choisi *Y* la variable à expliquer telle que :




Ainsi nous savons quelle équation nous devons retrouver en effectuant une régression linéaire et nous pourrons facilement juger de la qualité des régressions. Par la suite, nous changerons l'une des données obtenues par une valeur plus grande, et nous aurons ainsi simulé un exemple de données avec une valeur atypique.

#### Analyse des résultats

*Etape 1* (voir page 151)

Mettons nous dans le cas où nous ne connaissons pas la relation qui lie *Y* et *X*. On effectue alors une régression linéaire simple. On s'aperçoit que le coefficient de détermination $R^2$ est bon, alors que les graphes et les résultats indiquent clairement que beaucoup des hypothèses ne sont pas vérifiées. L'allure des résidus montre de façon évidente qu'on doit effectuer une régression polynomiale.

*Etape 2* (voir page 154)

On effectue une régression polynomiale de degré 2. On retrouve bien des résultats proches de la réalité. Ainsi, nous savons que cette régression est la meilleure que l'on puisse faire et nous pourrons nous en servir comme référence pour mettre en évidence la mauvaise qualité du modèle simple ou du modèle degré deux avec valeur atypique..

*Etape 3* (voir page 157)

On change une des données de façon à obtenir un échantillon avec une valeur atypique pour la 5ème observation. Les graphes confirment bien la présence d'une valeur atypique. On s'aperçoit alors que la présence de cette valeur modifie légèrement l'estimation de L'écart-type résiduel (il est un peu moins proche de la vraie valeur) et l'incertitude sur les prédictions.

*Etape 4* (voir page 160)

On refait la même expérience que dans l'étape 4 mais en mettant cette fois ci une valeur encore plus grande. On se rend alors vraiment compte qu'une valeur atypique conduit à des résultats faux. L'estimation de L'écart-type résiduel devient alors 5 fois plus grand que celui qui devrait être trouvé, les prédictions et leurs incertitudes deviennent donc mauvaises. Il est toujours intéressant lorsqu'on détecte une ou plusieurs valeurs atypiques, de faire une régression sans ces valeurs pour voir si elles changent de façon importante les résultats.

# AJUSTEMENT PAR REGRESSION SIMPLE

| No.obs | Y | | X | | | |
|---|---|---|---|---|---|---|
| 1 | 3.436624 | 1 | 1 | | | |
| 2 | 7.360805 | 1 | 2 | | | |
| 3 | 10.95506 | 1 | 3 | | | |
| 4 | 16.69787 | 1 | 4 | | | |
| 5 | 24.46581 | 1 | 5 | | | |
| 6 | 30.93739 | 1 | 6 | | | |
| 7 | 38.867 | 1 | 7 | | | |
| 8 | 49.78399 | 1 | 8 | | | |
| 9 | 59.36871 | 1 | 9 | | | |
| 10 | 71.04491 | 1 | 10 | | | |
| 11 | 84.02521 | 1 | 11 | | | |
| 12 | 96.93948 | 1 | 12 | | | |
| 13 | 111.1324 | 1 | 13 | | | |
| 14 | 128.0678 | 1 | 14 | | | |
| 15 | 143.509 | 1 | 15 | | | |
| 16 | 161.7631 | 1 | 16 | | | |
| 17 | 178.9092 | 1 | 17 | | | |
| 18 | 198.6647 | 1 | 18 | | | |
| 19 | 218.6694 | 1 | 19 | | | |
| 20 | 241.5801 | 1 | 20 | | | |
| 21 | 263.9548 | 1 | 21 | | | |
| 22 | 287.821 | 1 | 22 | | | |
| 23 | 311.4696 | 1 | 23 | | | |
| 24 | 336.706 | 1 | 24 | | | |
| 25 | 363.9353 | 1 | 25 | | | |
| 26 | 391.2174 | 1 | 26 | | | |
| 27 | 419.4449 | 1 | 27 | | | |
| 28 | 448.7912 | 1 | 28 | | | |
| 29 | 479.3748 | 1 | 29 | | | |
| 30 | 510.6553 | 1 | 30 | | | |
| 31 | 543.6711 | 1 | 31 | | | |
| 32 | 576.4625 | 1 | 32 | | | |
| 33 | 611.9016 | 1 | 33 | | | |
| 34 | 646.6932 | 1 | 34 | | | |
| 35 | 683.1719 | 1 | 35 | | | |
| 36 | 720.8871 | 1 | 36 | | | |
| 37 | 760.1786 | 1 | 37 | | | |
| 38 | 798.3278 | 1 | 38 | | | |
| 39 | 839.1883 | 1 | 39 | | | |
| 40 | 881.5113 | 1 | 40 | | | |
| 41 | 923.5883 | 1 | 41 | | | |
| 42 | 966.3947 | 1 | 42 | | | |
| 43 | 1011.538 | 1 | 43 | | | |
| 44 | 1056.314 | 1 | 44 | | | |
| 45 | 1103.717 | 1 | 45 | | | |
| 46 | 1151.238 | 1 | 46 | | | |
| 47 | 1199.29 | 1 | 47 | | | |
| 48 | 1249.429 | 1 | 48 | | | |
| 49 | 1298.461 | 1 | 49 | | | |
| 50 | 1350.74 | 1 | 50 | | | |

## Multiple Regression Results

régression simple

| | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | Intercept | X | | | | | | | | | |
| *b* | -219.84 | 27.4935 | | | | | | | | | |
| *s(b)* | 27.2629 | 0.93047 | | | | | | | | | |
| *t* | -8.0637 | 29.548 | | | | | | | | | |
| *p-value* | 0.0000 | 0.0000 | | | | | | | | | |
| VIF | | #REF! | | | | | | | | | |

### Table ANOVA

| Source | SS | df | MS | *F* | *Fcritiqu* | *p* -value |
|---|---|---|---|---|---|---|
| Regn. | 7870740 | 1 | 8E+06 | 873.08 | 4.04265 | 0.0000 |
| Erreur | 432714 | 48 | 9014.9 | | | |
| Total | 8303454 | 49 | | | | |

*s* 94.947

$R^2$ 0.9479

$R^{2\ (ajusté)}$ 0.9468

### Intervalles de prédiction (I.P)

X donné

| X |
|---|
| 2 |

| 1-α | (1-α) I.P de Y/X | | |
|---|---|---|---|
| 95% | -164.85 | + ou - | 197.75 |

| 1-α | (1-α) I.P de E[Y \| X] | | |
|---|---|---|---|
| 95% | -164.85 | + ou - | 51.592 |

### Matrice de covariance des coefficients de la régression

| | b0 | b1 |
|---|---|---|
| b0 | 743.267 | -22.077 |
| b1 | -22.077 | 0.86577 |

### Tableau récapitulatif sur les incertitudes de prédiction

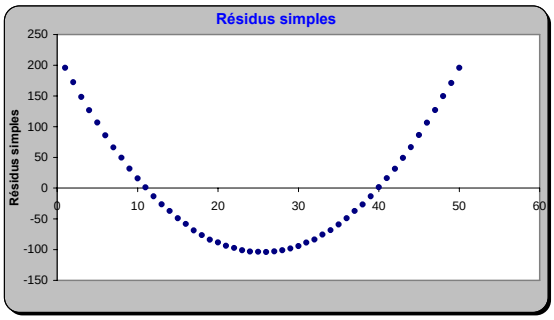
| | Estimateur | incertitude élargie + ou - U (k=2) | |
|---|---|---|---|
| prédiction sur un point (Y/X) | -164.8525881 | + ou - | 196.7058246 |
| prédiction sur la droite (E(Y/X)) | -164.8525881 | + ou - | 51.31944249 |

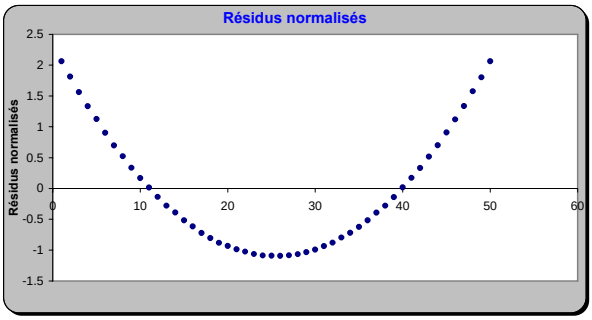
régression simple

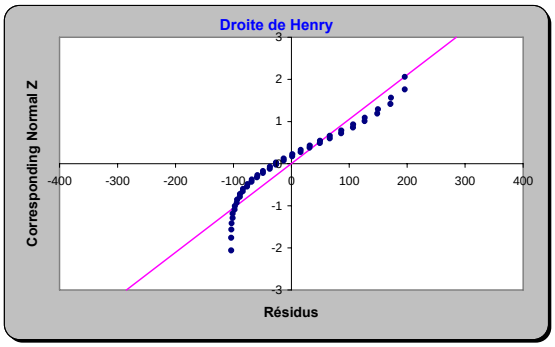
Choisir l'axe des abscisses pour les graphes des résidus

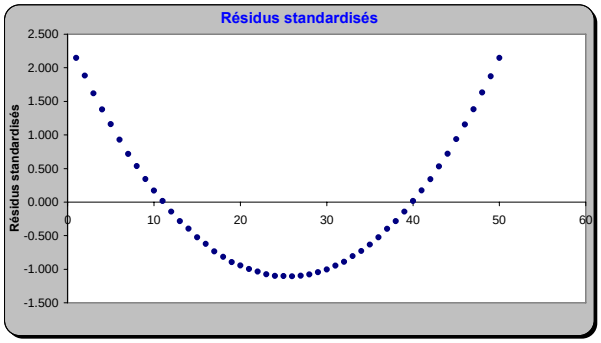
x

Durbin-Watson d = 0.022677

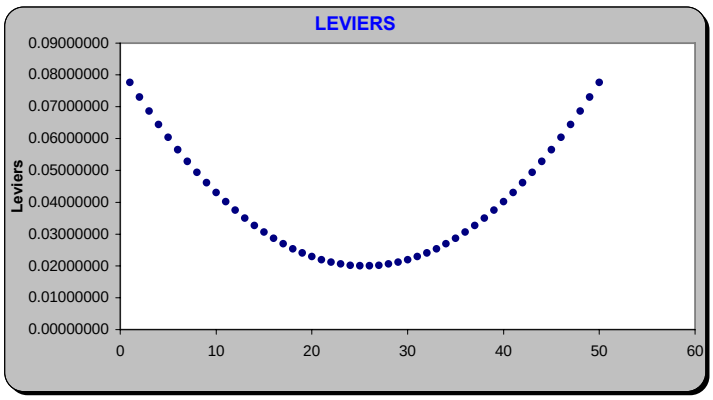


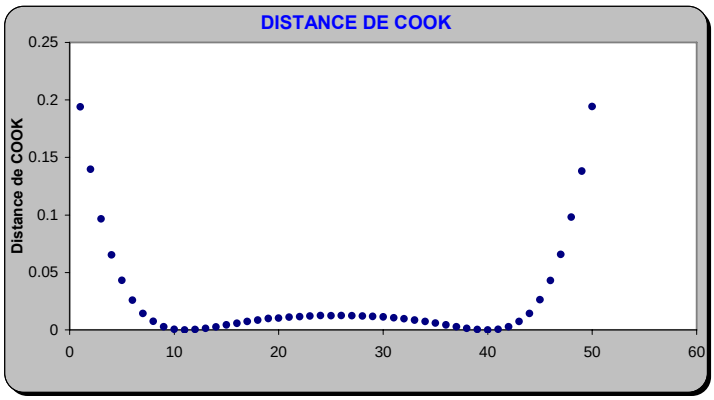




| Levier suspect pour une valeur supérieure à | 0.08 |
|---|---|




| Distance de Cook suspecte pour une valeur | 0.08333333 |
|---|---|



# AJUSTEMENT PAR REGRESSION POLYNOMIALE DE DEGRE DEUX

| No.obs | Y | | X | X2 | | |
|---|---|---|---|---|---|---|
| 1 | 3.436624 | 1 | 1 | 1 | | |
| 2 | 7.360805 | 1 | 2 | 4 | | |
| 3 | 10.95506 | 1 | 3 | 9 | | |
| 4 | 16.69787 | 1 | 4 | 16 | | |
| 5 | 24.46581 | 1 | 5 | 25 | | |
| 6 | 30.93739 | 1 | 6 | 36 | | |
| 7 | 38.867 | 1 | 7 | 49 | | |
| 8 | 49.78399 | 1 | 8 | 64 | | |
| 9 | 59.36871 | 1 | 9 | 81 | | |
| 10 | 71.04491 | 1 | 10 | 100 | | |
| 11 | 84.02521 | 1 | 11 | 121 | | |
| 12 | 96.93948 | 1 | 12 | 144 | | |
| 13 | 111.1324 | 1 | 13 | 169 | | |
| 14 | 128.0678 | 1 | 14 | 196 | | |
| 15 | 143.509 | 1 | 15 | 225 | | |
| 16 | 161.7631 | 1 | 16 | 256 | | |
| 17 | 178.9092 | 1 | 17 | 289 | | |
| 18 | 198.6647 | 1 | 18 | 324 | | |
| 19 | 218.6694 | 1 | 19 | 361 | | |
| 20 | 241.5801 | 1 | 20 | 400 | | |
| 21 | 263.9548 | 1 | 21 | 441 | | |
| 22 | 287.821 | 1 | 22 | 484 | | |
| 23 | 311.4696 | 1 | 23 | 529 | | |
| 24 | 336.706 | 1 | 24 | 576 | | |
| 25 | 363.9353 | 1 | 25 | 625 | | |
| 26 | 391.2174 | 1 | 26 | 676 | | |
| 27 | 419.4449 | 1 | 27 | 729 | | |
| 28 | 448.7912 | 1 | 28 | 784 | | |
| 29 | 479.3748 | 1 | 29 | 841 | | |
| 30 | 510.6553 | 1 | 30 | 900 | | |
| 31 | 543.6711 | 1 | 31 | 961 | | |
| 32 | 576.4625 | 1 | 32 | 1024 | | |
| 33 | 611.9016 | 1 | 33 | 1089 | | |
| 34 | 646.6932 | 1 | 34 | 1156 | | |
| 35 | 683.1719 | 1 | 35 | 1225 | | |
| 36 | 720.8871 | 1 | 36 | 1296 | | |
| 37 | 760.1786 | 1 | 37 | 1369 | | |
| 38 | 798.3278 | 1 | 38 | 1444 | | |
| 39 | 839.1883 | 1 | 39 | 1521 | | |
| 40 | 881.5113 | 1 | 40 | 1600 | | |
| 41 | 923.5883 | 1 | 41 | 1681 | | |
| 42 | 966.3947 | 1 | 42 | 1764 | | |
| 43 | 1011.538 | 1 | 43 | 1849 | | |
| 44 | 1056.314 | 1 | 44 | 1936 | | |
| 45 | 1103.717 | 1 | 45 | 2025 | | |
| 46 | 1151.238 | 1 | 46 | 2116 | | |
| 47 | 1199.29 | 1 | 47 | 2209 | | |
| 48 | 1249.429 | 1 | 48 | 2304 | | |
| 49 | 1298.461 | 1 | 49 | 2401 | | |
| 50 | 1350.74 | 1 | 50 | 2500 | | |

## Multiple Regression Results

Degré 2

| | 0 | 1 | 2 |
|---|---|---|---|
| | Intercept | X | X2 |
| *b* | 1.0437 | 2.00698 | 0.4997 |
| *s(b)* | 0.21602 | 0.01954 | 0.0004 |
| *t* | 4.83138 | 102.708 | 1345.4 |
| *p-value* | 0.0000 | 0.0000 | 0.0000 |
| VIF | | 16.6310 | 16.6310 |

### Table ANOVA

| Source | SS | df | MS | *F* | *Fcritiqu* | *p* -value |
|---|---|---|---|---|---|---|
| Regn. | 8303443 | 2 | 4E+06 | 2E+07 | 3.19505 | 0.0000 |
| Erreur | 11.2361 | 47 | 0.2391 | | | |
| Total | 8303454 | 49 | | | | |

*s* 0.4889

$R^2$ 1.0000

$R^{2\ (ajuste)}$ 1

### Intervalles de prédiction (I.P)

X donné

| X | X2 |
|---|---|
| 2 | 4 |

| 1-α | (1-α) I.P de Y/X |
|---|---|
| 95% | 7.0566 + ou - 1.0508 |

| 1-α | (1-α) I.P de E[Y \| X] |
|---|---|
| 95% | 7.0566 + ou - 0.3698 |

### Matrice de covariance des coefficients de la régression

| | b0 | b1 | b2 |
|---|---|---|---|
| b0 | 0.04667 | -0.0037 | 6E-05 |
| b1 | -0.0037 | 0.00038 | -7E-06 |
| b2 | 6.1E-05 | -7E-06 | 1E-07 |

### Tableau récapitulatif sur les incertitudes de prédiction

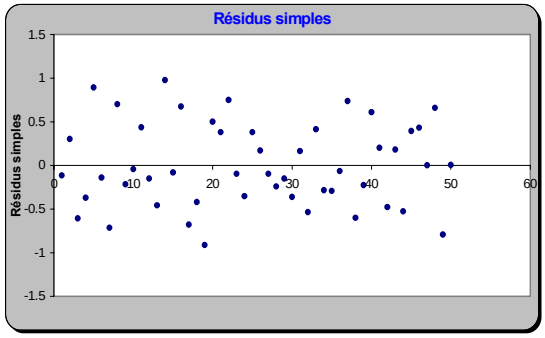
| | Estimateur | incertitude élargie + ou - U (k=2) |
|---|---|---|
| prédiction sur un point (Y/X) | 7.056598959 | + ou - 1.044709352 |
| prédiction sur la droite (E(Y/X)) | 7.056598959 | + ou - 0.367632743 |

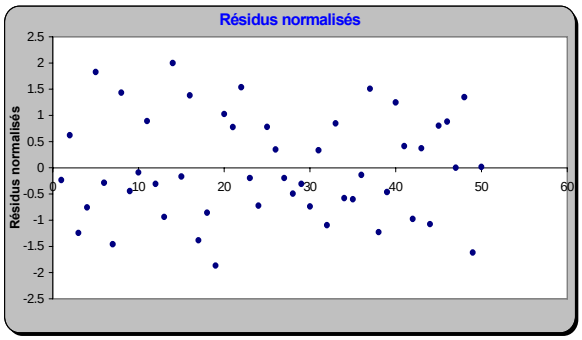
Degré 2

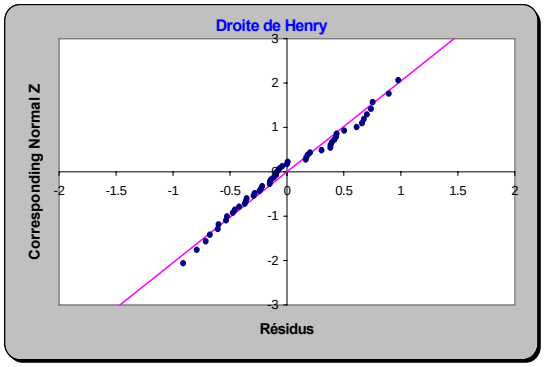
Choisir l'axe des abscisses pour les graphes des résidus

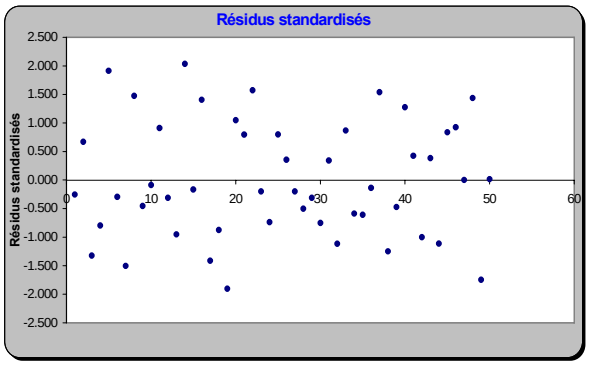
x

Durbin-Watson d = 2.526986

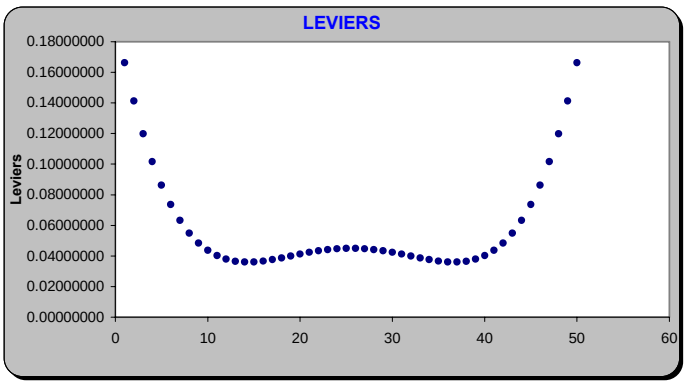


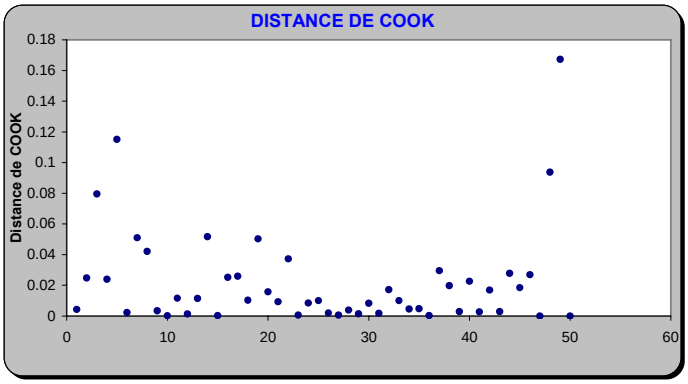




| Levier suspect pour une valeur supérieure à | 0.12 |
|---|---|




| Distance de Cook suspecte pour une valeur | 0.08510638 |
|---|---|



## AJUSTEMENT PAR REGRESSION POLYNOMIALE DEGRE DEUX
## données avec valeur atypique

| No.obs | Y | | X | X2 |
|---|---|---|---|---|
| 1 | 3.436624 | 1 | 1 | 1 |
| 2 | 7.360805 | 1 | 2 | 4 |
| 3 | 10.95506 | 1 | 3 | 9 |
| 4 | 20 | 1 | 4 | 16 |
| 5 | 24.46581 | 1 | 5 | 25 |
| 6 | 30.93739 | 1 | 6 | 36 |
| 7 | 38.867 | 1 | 7 | 49 |
| 8 | 49.78399 | 1 | 8 | 64 |
| 9 | 59.36871 | 1 | 9 | 81 |
| 10 | 71.04491 | 1 | 10 | 100 |
| 11 | 84.02521 | 1 | 11 | 121 |
| 12 | 96.93948 | 1 | 12 | 144 |
| 13 | 111.1324 | 1 | 13 | 169 |
| 14 | 128.0678 | 1 | 14 | 196 |
| 15 | 143.509 | 1 | 15 | 225 |
| 16 | 161.7631 | 1 | 16 | 256 |
| 17 | 178.9092 | 1 | 17 | 289 |
| 18 | 198.6647 | 1 | 18 | 324 |
| 19 | 218.6694 | 1 | 19 | 361 |
| 20 | 241.5801 | 1 | 20 | 400 |
| 21 | 263.9548 | 1 | 21 | 441 |
| 22 | 287.821 | 1 | 22 | 484 |
| 23 | 311.4696 | 1 | 23 | 529 |
| 24 | 336.706 | 1 | 24 | 576 |
| 25 | 363.9353 | 1 | 25 | 625 |
| 26 | 391.2174 | 1 | 26 | 676 |
| 27 | 419.4449 | 1 | 27 | 729 |
| 28 | 448.7912 | 1 | 28 | 784 |
| 29 | 479.3748 | 1 | 29 | 841 |
| 30 | 510.6553 | 1 | 30 | 900 |
| 31 | 543.6711 | 1 | 31 | 961 |
| 32 | 576.4625 | 1 | 32 | 1024 |
| 33 | 611.9016 | 1 | 33 | 1089 |
| 34 | 646.6932 | 1 | 34 | 1156 |
| 35 | 683.1719 | 1 | 35 | 1225 |
| 36 | 720.8871 | 1 | 36 | 1296 |
| 37 | 760.1786 | 1 | 37 | 1369 |
| 38 | 798.3278 | 1 | 38 | 1444 |
| 39 | 839.1883 | 1 | 39 | 1521 |
| 40 | 881.5113 | 1 | 40 | 1600 |
| 41 | 923.5883 | 1 | 41 | 1681 |
| 42 | 966.3947 | 1 | 42 | 1764 |
| 43 | 1011.538 | 1 | 43 | 1849 |
| 44 | 1056.314 | 1 | 44 | 1936 |
| 45 | 1103.717 | 1 | 45 | 2025 |
| 46 | 1151.238 | 1 | 46 | 2116 |
| 47 | 1199.29 | 1 | 47 | 2209 |
| 48 | 1249.429 | 1 | 48 | 2304 |
| 49 | 1298.461 | 1 | 49 | 2401 |
| 50 | 1350.74 | 1 | 50 | 2500 |

## Multiple Regression Results

degré 2 avec valeur atypique

| | 0 | 1 | 2 |
|---|---|---|---|
| | Intercept | X | X2 |
| *b* | 1.49757 | 1.97547 | 0.5002 |
| *s(b)* | 0.27788 | 0.02514 | 0.0005 |
| *t* | 5.38932 | 78.5923 | 1046.9 |
| *p-value* | 0.0000 | 0.0000 | 0.0000 |
| VIF | | 16.6310 | 16.6310 |

### Table ANOVA

| Source | SS | df | MS | *F* | *Fcritiqu* | *p* -value |
|---|---|---|---|---|---|---|
| Regn. | 8300378 | 2 | 4E+06 | 1E+07 | 3.19505 | 0.0000 |
| Erreur | 18.5916 | 47 | 0.3956 | | | |
| Total | 8300397 | 49 | | | | |

*s* 0.6289

$R^2$ 1.0000 $R^{2\ (ajuste)}$ 1

### Intervalles de prédiction (I.P)

| X donné | X | X2 |
|---|---|---|
| | 2 | 4 |

| 1-α | (1-α) I.P de Y/X |
|---|---|
| 95% | 7.4494 + ou - 1.3517 |

| 1-α | (1-α) I.P de E[Y \| X] |
|---|---|
| 95% | 7.4494 + ou - 0.4757 |

### Matrice de covariance des coefficients de la régression

| | b0 | b1 | b2 |
|---|---|---|---|
| b0 | 0.07722 | -0.0061 | 0.0001 |
| b1 | -0.0061 | 0.00063 | -1E-05 |
| b2 | 0.0001 | -1E-05 | 2E-07 |

### Tableau récapitulatif sur les incertitudes de prédiction

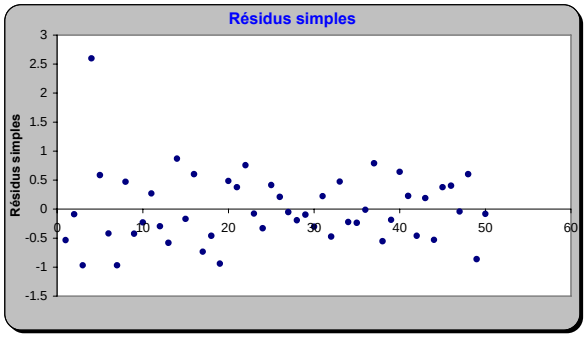
| | Estimateur | incertitude élargie + ou - U (k=2) |
|---|---|---|
| prédiction sur un point (Y/X) | 7.449396559 | + ou - 1.343834873 |
| prédiction sur la droite (E(Y/X)) | 7.449396559 | + ou - 0.472894877 |

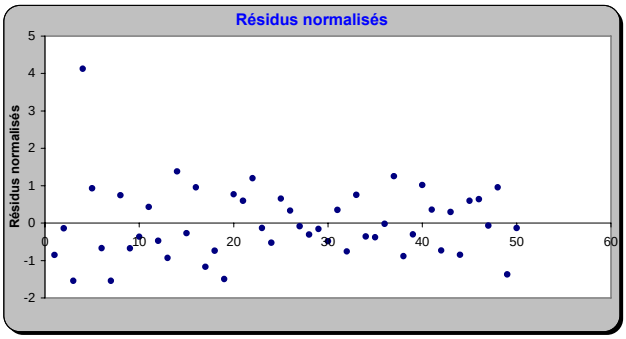
degré 2 avec valeur atypique

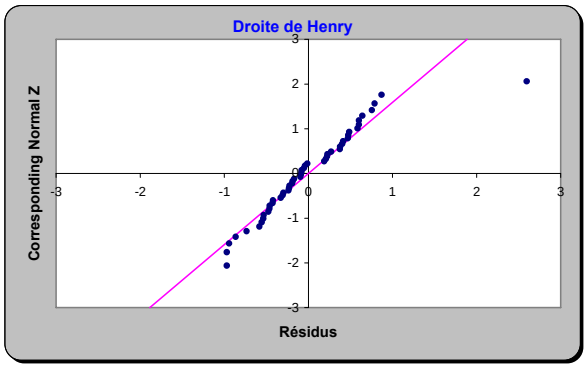
Choisir l'axe des abscisses pour les graphes des résidus

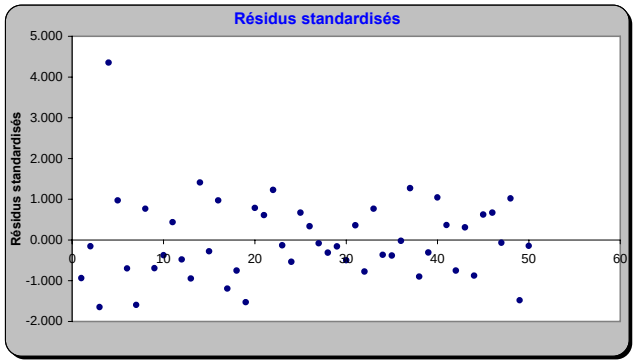
x

Durbin-Watson d = 2.337217

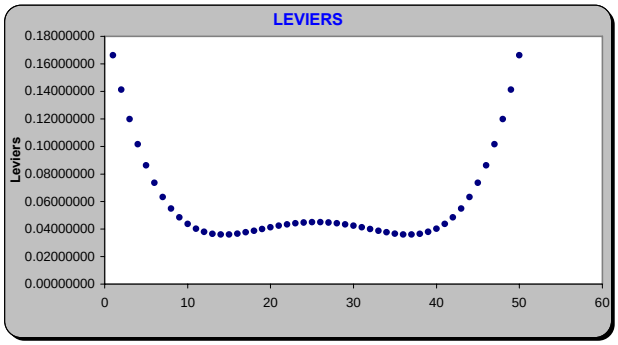


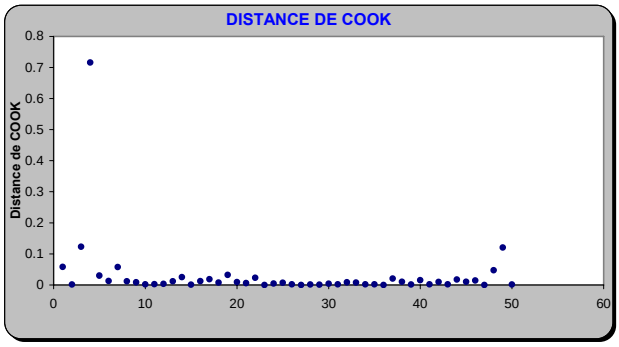




| Levier suspect pour une valeur supérieure à | 0.12 |
|---|---|




| Distance de Cook suspecte pour une valeur | 0.08510638 |
|---|---|



## AJUSTEMENT PAR REGRESSION POLYNOMIALE DEGRE DEUX
## données avec grande valeur atypique

| No.obs | Y | | X | X2 | | |
|---|---|---|---|---|---|---|
| 1 | 3.436624 | 1 | 1 | 1 | | |
| 2 | 7.360805 | 1 | 2 | 4 | | |
| 3 | 10.95506 | 1 | 3 | 9 | | |
| 4 | 35 | 1 | 4 | 16 | | |
| 5 | 24.46581 | 1 | 5 | 25 | | |
| 6 | 30.93739 | 1 | 6 | 36 | | |
| 7 | 38.867 | 1 | 7 | 49 | | |
| 8 | 49.78399 | 1 | 8 | 64 | | |
| 9 | 59.36871 | 1 | 9 | 81 | | |
| 10 | 71.04491 | 1 | 10 | 100 | | |
| 11 | 84.02521 | 1 | 11 | 121 | | |
| 12 | 96.93948 | 1 | 12 | 144 | | |
| 13 | 111.1324 | 1 | 13 | 169 | | |
| 14 | 128.0678 | 1 | 14 | 196 | | |
| 15 | 143.509 | 1 | 15 | 225 | | |
| 16 | 161.7631 | 1 | 16 | 256 | | |
| 17 | 178.9092 | 1 | 17 | 289 | | |
| 18 | 198.6647 | 1 | 18 | 324 | | |
| 19 | 218.6694 | 1 | 19 | 361 | | |
| 20 | 241.5801 | 1 | 20 | 400 | | |
| 21 | 263.9548 | 1 | 21 | 441 | | |
| 22 | 287.821 | 1 | 22 | 484 | | |
| 23 | 311.4696 | 1 | 23 | 529 | | |
| 24 | 336.706 | 1 | 24 | 576 | | |
| 25 | 363.9353 | 1 | 25 | 625 | | |
| 26 | 391.2174 | 1 | 26 | 676 | | |
| 27 | 419.4449 | 1 | 27 | 729 | | |
| 28 | 448.7912 | 1 | 28 | 784 | | |
| 29 | 479.3748 | 1 | 29 | 841 | | |
| 30 | 510.6553 | 1 | 30 | 900 | | |
| 31 | 543.6711 | 1 | 31 | 961 | | |
| 32 | 576.4625 | 1 | 32 | 1024 | | |
| 33 | 611.9016 | 1 | 33 | 1089 | | |
| 34 | 646.6932 | 1 | 34 | 1156 | | |
| 35 | 683.1719 | 1 | 35 | 1225 | | |
| 36 | 720.8871 | 1 | 36 | 1296 | | |
| 37 | 760.1786 | 1 | 37 | 1369 | | |
| 38 | 798.3278 | 1 | 38 | 1444 | | |
| 39 | 839.1883 | 1 | 39 | 1521 | | |
| 40 | 881.5113 | 1 | 40 | 1600 | | |
| 41 | 923.5883 | 1 | 41 | 1681 | | |
| 42 | 966.3947 | 1 | 42 | 1764 | | |
| 43 | 1011.538 | 1 | 43 | 1849 | | |
| 44 | 1056.314 | 1 | 44 | 1936 | | |
| 45 | 1103.717 | 1 | 45 | 2025 | | |
| 46 | 1151.238 | 1 | 46 | 2116 | | |
| 47 | 1199.29 | 1 | 47 | 2209 | | |
| 48 | 1249.429 | 1 | 48 | 2304 | | |
| 49 | 1298.461 | 1 | 49 | 2401 | | |
| 50 | 1350.74 | 1 | 50 | 2500 | | |

## Multiple Regression Results

degré 2 avec grande valeur atypique

| | 0 | 1 | 2 |
|---|---|---|---|
| | Intercept | X | X2 |
| *b* | 3.55931 | 1.83235 | 0.5024 |
| *s(b)* | 1.11369 | 0.10074 | 0.0019 |
| *t* | 3.19594 | 18.1889 | 262.36 |
| *p-value* | 0.0025 | 0.0000 | 0.0000 |
| VIF | | 16.6310 | 16.6310 |

### Table ANOVA

| Source | SS | df | MS | *F* | *Fcritiqu* | *p* -value |
|---|---|---|---|---|---|---|
| Regn. | 8286479 | 2 | 4E+06 | 652073 | 3.19505 | 0.0000 |
| Erreur | 298.635 | 47 | 6.3539 | | | |
| Total | 8286778 | 49 | | | | |

*s* 2.5207

$R^2$ 1.0000

$R^{2\ (ajusté)}$ 1

### Intervalles de prédiction (I.P)

X donné

| X | X2 |
|---|---|
| 2 | 4 |

| 1-α | (1-α) I.P de Y/X |
|---|---|
| 95% | 9.23369 + ou - 5.4175 |

| 1-α | (1-α) I.P de E[Y \| X] |
|---|---|
| 95% | 9.2337 + ou - 1.9064 |

### Matrice de covariance des coefficients de la régression

| | b0 | b1 | b2 |
|---|---|---|---|
| b0 | 1.24032 | -0.0982 | 0.0016 |
| b1 | -0.0982 | 0.01015 | -0.0002 |
| b2 | 0.00162 | -0.0002 | 4E-06 |

### Tableau récapitulatif sur les incertitudes de prédiction

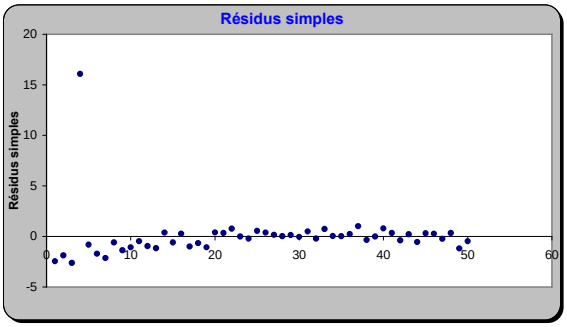
| | Estimateur | incertitude élargie + ou - U (k=2) |
|---|---|---|
| prédiction sur un point (Y/X) | 9.233688737 | + ou - 5.385901272 |
| prédiction sur la droite (E(Y/X)) | 9.233688737 | + ou - 1.895296194 |

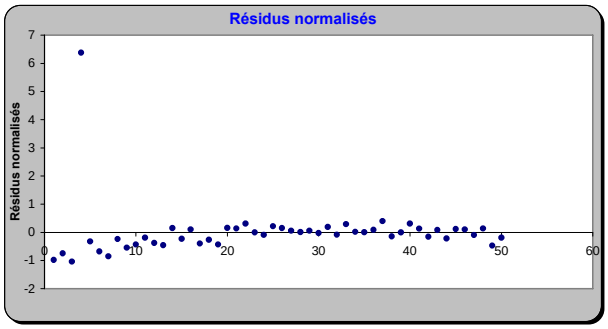
degré 2 avec grande valeur atypi

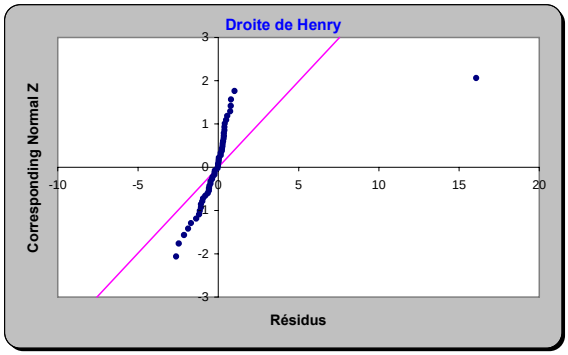
Choisir l'axe des abscisses pour les graphes des résidus

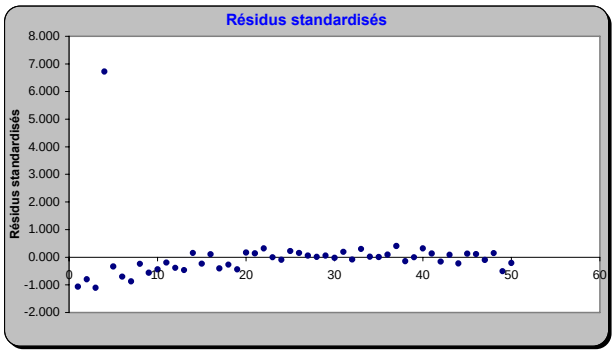
x

Durbin-Watson d = 2.214648

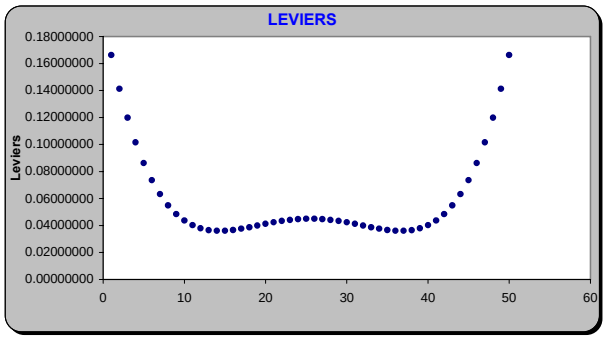


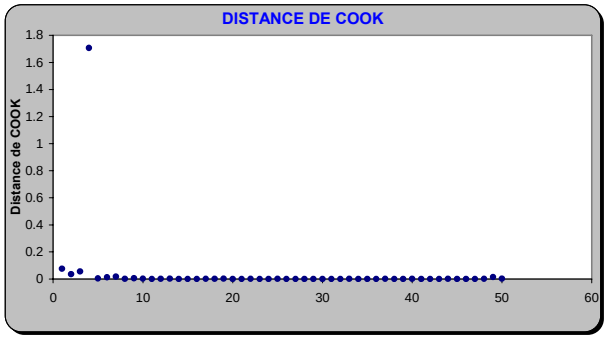




| Levier suspect pour une valeur supérieure à | 0.12 |
|---|---|




| Distance de Cook suspecte pour une valeur | 0.08510638 |
|---|---|



# BIBLIOGRAPHIE


[1] Besse P. *Pratique de la modélisation stochastique*. Available via http://www.lsp.upstlse.fr/Besse/enseignement.html.

[2] CERESTA (Centre d'Enseignement et de Recherche Statistique Appliquée), *Aide-mémoire pratique des techniques statistiques pour ingénieurs et techniciens supérieurs*. Revue de Statistique Appliquée, vol. XXXIV, numéro spécial, 1986.

[3] Cetama, *Statistique appliquée à l'exploitation des mesures*, Masson 2e édition, 1986.

[4] Confais, J. et M. Le Guen, *La régression linéaire sous SAS*, document de travail de la Direction des Statistiques Démographiques et Sociales de l'INSEE n° F 9605, 3ème édition en 2003.

[5] Dagnelie P. *Théorie et méthodes statistiques* (2e éd). Agronomiques Gembloux 1998.

[6] Dodge Y. *Statistiques, dictionnaire encyclopédique*. Dunod, 1993.

[7] Draper NR, Smith H. *Applied Regression Analysis* (3rd edition). New York: Wiley 1998.

[8] Gaudoin O. *Principes et méthodes statistiques*. Available via http://www-ljk.imag.fr/membres/Olivier.Gaudoin/.

[9] Johnston J. *Econometric Methods*,. New York: McGraw Hill, 1971.

[10] Lebart L., Morineau A. and Fénelon J-P. *Traitement des données statistiques : méthodes et programmes*. Paris : Dunod , 1984.

[11] Steppan D., Werner J. and Yeater R. *Essential regression and experimental design for chemists and engineers*, 1998. Available via http://www.jowerner.homepage.tonline.de/ERPref.html.

[12] Tenenhaus, M. *Méthodes statistiques en gestion*, Dunod Entreprise, 1994.

[13] Tomassone R., Lesquoy E., Millier R. *La régression, nouveaux regards sur une ancienne méthode statistique*, Masson, Paris, 1983.