Chapitre 1 : REGRESSION LINÉAIRE SIMPLE

Plan du Chapitre :

I.- Introduction :

En analyse de régression, on cherche à expliquer une variable Y qui dépend d'une ou de plusieurs variables explicatives X_1, X_2, …,X_k.

Pour cela, un modèle peut représenter la relation existante entre Y et les X_i. Ce modèle servira aussi pour faire des prévisions :

$$Y = f(X_1, X_2, \cdots, X_k) + \varepsilon$$

La variable Y s'appelle la variable expliquée, dépendante ou endogène. Alors que les variables X_i s'appellent explicatives, indépendantes ou exogènes. Il est nécessaire de préciser la nature de la fonction f : fonction linéaire, affine, exponentielle,…Dans le cas où : $f(X)=aX+b$, le modèle est dit linéaire, avec a et b sont des constantes inconnues.

Le facteur ε est de nature aléatoire et il suit une loi de probabilité. ε appelé bruit blanc ou facteur de perturbation. Il représente les erreurs sur les observations de Y.

S'appuyant sur des observations y_1, y_2,…,y_n de la variable Y, l'analyse de régression consiste à élaborer un modèle explicatif qui sera analysé statistiquement par l'estimation de ses paramètres et par divers tests d'hypothèses.

II.- Présentation du modèle :

Lorsqu'il n'y a qu'une seule variable explicative, on dira que le modèle de régression est simple. Le modèle de régression linéaire simple s'écrit :

$$Y = b + aX + \varepsilon,$$

où Y est la variable à expliquer, variable dépendante ou variable endogène.

X est la variable explicative, variable indépendante ou variable exogène.

a et b son les paramètres du modèle.

a est le coefficient de régression, c'est la pente de la droite (il mesure la variation de *Y* lorsque *X* augmente d'une unité).

b est la valeur prise par *Y* lorsque *X*=0.

ε est le facteur erreur aléatoire, non observable. Il comprend les erreurs de mesure et les autres facteurs explicatifs non pris en compte.

III.- Hypothèses sur le modèle :

D'abord, on suppose l'existence d'une relation linéaire entre Y et X.

La variable explicative X est mesurée sans erreur.

Aussi, on suppose l'absence des erreurs de spécification, c'est-à-dire que toutes les variables X qui sont importantes ou principales pour l'explication de la variable Y sont inclus dans la définition du modèle.

Les erreurs aléatoires suivent une loi de probabilité normale de moyenne nulle et de variance constante : $\varepsilon \to N(0,\sigma_\varepsilon)$.

Le fait de supposer que σ_ε^2 est constante, s'appelle hypothèse de « Homoscédasticité ».

Comme ε est aléatoire, alors Y est aussi une variable aléatoire. Alors qu'il n'est pas nécessaire que la variable X soit aussi aléatoire.

$$E(y_i) = E(b + ax_i + \varepsilon_i) = b + ax_i + E(\varepsilon_i) \Rightarrow E(y_i) = b + ax_i$$

$$\sigma_Y^2 = E[y_i - E(y_i)]^2 = E[a + bx_i + \varepsilon_i - a - bx_i]^2$$

$$= E[\varepsilon_i^2] = \sigma_\varepsilon^2$$

$$\Rightarrow y_i \to N(E(y_i); \sigma_\varepsilon^2)$$

Les erreurs aléatoires ε sont non corrélées avec *X*.

Les erreurs ε sont non corrélées entre elles. C'est-à-dire $Cov(\varepsilon_i, \varepsilon_j) = 0; \quad \forall i \neq j$ cette hypothèse s'appelle de non-auto corrélation des erreurs.

IV.- Estimation des paramètres :

La méthode que l'on utilise pour estimer les paramètres d'un modèle de régression est la méthode des moindres carrées, bien que l'on peut aussi utiliser la méthode de maximum de vraisemblance.. Soient, donc, *n* observations de la variable *Y* : $y_1, y_2, \ldots, y_n$. D'où, on peut écrire le modèle :

$$y_i = b + ax_i + \varepsilon ; \qquad i = 1,2,\ldots,n$$

Le modèle estimé à partir des n observations sera :

$$\hat{y}_i = \hat{b} + \hat{a}x_i$$

Il s'agit de trouver $\hat{a}$ et $\hat{b}$ telle que si on définit le i^{ème} résidu $e_i = y_i - \hat{y}_i$, la méthode des moindres carrées consiste à minimiser la somme des carrées des résidus

$$\text{Min} \sum_{i=1}^{n} e_i^2 \equiv \text{Min} \sum_{i=1}^{n} \left(y_i - \hat{y}_i\right)^2$$

On pose $e^2 = \sum_{i=1}^{n} \left(y_i - \hat{b} - \hat{a}x_i\right)^2$

Alors, $$\left.\begin{array}{l} \dfrac{\partial e^2}{\partial \hat{b}} = 0 \\ \dfrac{\partial e^2}{\partial \hat{a}} = 0 \end{array}\right\} \Rightarrow \begin{cases} \sum_{i=1}^{n} 2\left(y_i - \hat{b} - \hat{a}x_i\right)(-1) = 0 \\ \sum_{i=1}^{n} 2\left(y_i - \hat{b} - \hat{a}x_i\right)(-x_i) = 0 \end{cases}$$

$$\Rightarrow \begin{cases} -\sum_{i=1}^{n} y_i + n\hat{b} + \hat{a}\sum_{i=1}^{n} x_i = 0 \\ -\sum_{i=1}^{n} y_i x_i + \hat{b}\sum_{i=1}^{n} x_i + \hat{a}\sum_{i=1}^{n} x_i^2 + 0 \end{cases} \Rightarrow \begin{cases} n\hat{b} + \hat{a}\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i \quad (1) \\ \hat{b}\sum_{i=1}^{n} x_i + \hat{a}\sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i \quad (2) \end{cases}$$

$$(1) \Rightarrow \hat{b} = \frac{1}{n}\sum_i y_i - \hat{a}\frac{1}{n}\sum_i x_i \Rightarrow \boxed{\hat{b} = \overline{Y} - \hat{a}\overline{X}}$$

$$(2) \Rightarrow \left(\overline{Y} - \hat{a}\overline{X}\right)\sum_{i=1}^{n} x_i + \hat{a}\sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i \qquad \Rightarrow \overline{Y}\sum_{i=1}^{n} x_i - \hat{a}\overline{X}\sum_{i=1}^{n} x_i + \hat{a}\sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_j$$

$$\Rightarrow \boxed{\hat{a} = \frac{\sum_{i=1}^{n} x_i y_i - n\overline{X}\overline{Y}}{\sum_{i=1}^{n} x_i^2 - n\overline{X}^2} = \frac{\sum_{i=1}^{n} \left(x_i - \overline{X}\right)\left(y_i - \overline{Y}\right)}{\sum_{i=1}^{n} \left(x_i - \overline{X}\right)^2}}$$

Ces estimateurs sont les mêmes que ceux que l'on obtiendrait par la méthode de maximum de vraisemblances en supposant que les erreurs théoriques sont normalement distribuées.

V.- Caractéristiques des estimateurs :

Ces estimateurs sont des fonctions linéaires des observations $y_1, y_2, \cdots, y_n$. Ces estimateurs sont non-biaisés,

$E(\hat{a}) = a$ et $E(\hat{b}) = b$

en effet :

$$E(\hat{a}) = E\left[\frac{\sum_{i=1}^{n}(x_i - \bar{X})(y_i - \bar{Y})}{\sum_{i=1}^{n}(x_i - \bar{X})^2}\right] = \frac{1}{\sum_{i=1}^{n}(x_i - \bar{X})^2}\sum_{i=1}^{n}\left((x_i - \bar{X})\left[E(y_i) - E(\bar{Y})\right]\right)$$

$$= \frac{1}{\sum_{i=1}^{n}(x_i - \bar{X})^2}\sum_{i=1}^{n}\left((x_i - \bar{X})(b + ax_i - b - a\bar{X})\right) = a\frac{\sum(x_i - \bar{X})^2}{\sum(x_i - \bar{X})^2}$$

$$\Rightarrow E(\hat{a}) = a$$

et $E(\hat{b}) = E(\bar{Y} - a\bar{X}) = E(\bar{Y}) - \bar{X}E(a)$

Mais, $E(\bar{Y}) = E\left(\frac{\sum y_i}{n}\right) = \frac{1}{n}\sum[E(y_i)] = \frac{1}{n}\sum(b + ax_i) = b + a\bar{X}$, parce que $E(y_i) = b + ax_i$

Donc $E(\hat{b}) = b + a\bar{X} - a\bar{X} = b$

Les variances théoriques de ces estimateurs sont :

$$Var(\hat{a}) = \frac{\sigma_\varepsilon^2}{\sum_{i=1}^{n}(x_i - \bar{X})^2} \qquad \text{et} \qquad Var(\hat{b}) = \sigma_\varepsilon^2\left[\frac{1}{n} + \frac{\bar{X}^2}{\sum_{i=1}^{2}(x_i - \bar{X})^2}\right]$$

On peut démontrer que les estimateurs des moindres carrées sont des estimateurs linéaires non biaisés à variance minimale (c'est-à-dire efficace). On dit aussi qu'ils sont des estimateurs **BLUE** « **B**est **L**inear **U**nbiased **E**stimators ».

Connaissant la moyenne et la variance des estimateurs et ayant supposé que les erreurs, et donc les y_i, sont distribuées normalement, on peut conclure que les estimateurs $\hat{a}$ et $\hat{b}$, étant des fonctions linéaires des observations, obéissent eux-mêmes à une loi normale.

VI.- Intervalle de confiance sur les paramètres :

Les estimateurs $\hat{a}$ et $\hat{b}$ suivent une loi normale, parce qu'ils sont des fonctions linéaires des observations y_i, qui sont distribuées selon une loi normale.

On peut construire des intervalles de confiance pour chacun des paramètres *a* et *b*, pour cela, il faut au préalable estimer la variance des erreurs, σ_ε^2.

Les résidus e_i étant des estimateurs des erreurs théoriques ε_i, on doit se servir de la variance des résidus (notée S_e^2) comme estimateur de la variance des erreurs, la formule est donnée par :

$$\hat{\sigma}_\varepsilon^2 = S_e^2 = \frac{\sum_i e_i^2}{n-2} = \frac{\sum_i (y_i - \hat{y}_i)^2}{n-2} = \frac{\sum_i (y_i - b - ax_i)^2}{n-2}$$

On peut montrer que cet estimateur est non biaisé, c'est-à-dire : $E(\hat{\sigma}_\varepsilon^2) = \sigma_\varepsilon^2$.

Les estimateurs des paramètres de la droite de régression sont des statistiques qui suivent des lois de Student à (n-2) degrés de liberté :

$$\frac{\hat{b} - b}{S_e \sqrt{\frac{1}{n} + \frac{\overline{X}^2}{\sum_{i=1}^{n} (x_i - \overline{X})^2}}} \to t_{n-2} \qquad \text{et} \qquad \frac{\hat{a} - a}{S_e \sqrt{\frac{1}{\sum_{i=1}^{n} (x_i - \overline{X})^2}}} \to t_{n-2}$$

Les intervalles de confiance à un seuil α pour les paramètres *a* et *b* sont données par :

Pour *a* :
$$\left[\hat{a} - t_{1-\frac{\alpha}{2};n-2} S_e \sqrt{\frac{1}{\sum_{i=1}^{n} (x_i - \overline{X})^2}} \; ; \; \hat{a} + t_{1-\frac{\alpha}{2};n-2} S_e \sqrt{\frac{1}{\sum_{i=1}^{n} (x_i - \overline{X})^2}} \right]$$

Pour *b* :
$$\left[\hat{b} - t_{1-\frac{\alpha}{2};n-2} S_e \sqrt{\frac{1}{n} + \frac{\overline{X}^2}{\sum_{i=1}^{n} (x_i - \overline{X})^2}} \; ; \; \hat{b} + t_{1-\frac{\alpha}{2};n-2} S_e \sqrt{\frac{1}{n} + \frac{\overline{X}^2}{\sum_{i=1}^{n} (x_i - \overline{X})^2}} \right]$$

Comme σ_ε^2 était aussi inconnue et on l'a estimé, on peut aussi construire un intervalle de confiance pour cette variance.

On sait que $(n-2)\frac{\hat{\sigma}_\varepsilon^2}{\sigma_\varepsilon^2}$ suit une loi de χ^2 à (n-2) degrés de liberté. On part donc de $P\left(A < (n-2)\frac{\hat{\sigma}_\varepsilon^2}{\sigma_\varepsilon^2} < B \right) = 1 - \alpha$, l'intervalle de confiance pour σ_ε^2 est alors

$\left[\frac{(n-2)\hat{\sigma}_\varepsilon^2}{B} ; \frac{(n-2)\hat{\sigma}_\varepsilon^2}{A} \right]$, ou A et B sont déduit à partit de la table statistique de la loi χ^2 à (n-2) degrés de liberté.

VII.- Qualité de l'ajustement :

Plusieures tests sont utilisables pour déterminer la qualité de la représentation. Ces tests sont centrés autour de deux questions:

1) dans quelle mesure le phénomène est-il bien représenté par la droite qui vient d'être définie?
2) dans quelle mesure peut-on faire confiance aux coefficients b et a du modèle?

VII.-1.- Validité globale du modèle:

En construisant le modèle de régression nous avons supposé que *Y* dépendait de *X*. Il convient de tester cette hypothèse en la comparant avec l'hypothèse nulle selon laquelle *Y* est indépendant de *X*, c'est-à-dire que quelle que soit la valeur de *X*, nous obtenons toujours approximativement la même valeur de *Y*.

Avec l'hypothèse "Y dépend de X", nous obtenons des prévisions plus proches de la réalité. Il s'agit de voir si cette seconde hypothèse améliore suffisamment la prévision pour pouvoir rejeter l'hypothèse nulle.

a.- Lois des écarts:

La loi des écarts permet de relier l'erreur associée à l'hypothèse nulle et l'erreur associée à l'hypothèse "*Y* dépend de *X*".

L'erreur attachée à l'hypothèse nulle est mesurée par la *dispersion* totale des Y_i, c'est-à-dire par la somme des carrés des écarts des Y_i par rapport à la moyenne *Y*:

$$\textbf{Dipersion totale}=\sum_i \left(Y_i-\overline{Y}\right)^2$$

L'erreur attachée à la seconde hypothèse, ou encore dispersion résiduelle est donnée par e^2, somme des carrés des écarts entre les observations Y_i et les valeurs estimées $\hat{Y}_i$ par le modèle:

$$\textbf{Dispersion résiduelle}=\sum_i \left(\hat{Y}_i-Y_i\right)^2$$

La différence entre la dispersion totale et la dispersion résiduelle correspond à la dispersion expliquée par le modèle de régression, compte tenu du fait que

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

On en tire la décomposition suivante:

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

relation connue sous le nom de loi des écarts, nous pouvons écrire:

$$\textbf{Dispersion expliquée} = \Sigma(\hat{Y}_i-Y)^2$$

Donc on a:

Dispersion totale=Dispersion expliquée+Dispersion résiduelle.

b.- Coefficient de détermination et coefficient de corrélation:

Coefficient de détermination R^2:

Un premier indicateur de qualité de la représentation consiste à mettre en relation la dispersion expliquée par le modèle et la dispersion totale des données: le coefficient de détermination R^2 mesure le pouvoir explicatif du modèle en évaluant le pourcentage de l'information restituée par le modèle par rapport à la qualité d'information initiale:

$$R^2 = \frac{\sum_i \left(\hat{Y}_i - \bar{Y}\right)^2}{\sum_i \left(Y_i - \bar{Y}\right)^2} = \frac{\text{dispersion expliquée}}{\text{dispersion totale}}$$

Coefficient de corrélation linéaire R:

Le coefficient de corrélation est *R*, racine carré du coefficient de détermination. C'est l'indicateur le plus couramment employé.

Le coefficient de corrélation linéaire a pour objet de mesurer l'intensité de la liaison linéaire entre deux variables statistiques X et Y.

On peut le calculer à l'aide de plusieurs formules différentes.

En premier lieu, d'après la définition qui vient d'être donnée, nous avons:

$$R = \sqrt{\frac{\sum_i \left(\hat{Y}_i - \bar{Y}\right)^2}{\sum_i \left(Y_i - \bar{Y}\right)^2}}$$

On montre que R est obtenu également à l'aide des formules suivantes, où σ_X et σ_Y représentent les écarts-type respectives des X_i et des Yi:

$$R = \frac{E\left[\left(X - E(X)\right)\left(Y - E(Y)\right)\right]}{\sqrt{Var(X)Var(Y)}} = \frac{Cov(X,Y)}{\sigma_X \sigma_Y}$$

$$R = \frac{\sigma_{XY}}{\sigma_X \sigma_Y} \quad \text{et} \quad R = a\frac{\sigma_X}{\sigma_Y}$$

Racine carée de R^2, c'est-à-dire d'un chiffre au plus égal à 1, *R* a une valeur absolue également au plus égale à 1. Cette définition montre que le coefficient de corrélation possède le même signe que la covariance et qu'il est toujours compris entre -1 et 1. Donc $-1 \leq R \leq 1$.

Le signe du coefficient de corrélation linéaire indique le sens de la relation entre X et Y. *R* est positif (covariance ou coefficient de régression *a* positifs) ou négatif (cas inverse).

Un *R* très élevé en valeur absolue concrétise une relation étroite entre *X* et *Y*, croissante si *R* est positif et décroissante, si *R* est négatif.

R=1 : dans ce cas les points se trouvent tous sur une même droite croissante, on parle de correlation linéaire positive parfaite.

R=-1 : dans cas les points se trouvent tous sur une même droite décroisaante, on parle de corrélation linéaire négative parfaite.

R=0 : dans ce cas il n'y a aucune dépendance linéaire entre les deux variables, on parle de corrélation linéaire nulle.

Une valeur de *R* faible en termes absolus caractérise une absence de relation linéaire entre *X* et *Y*, mais pas nécessairement l'absence de liaison entre les variables.

c.- Analyse de la variance pour la régression (test *F*):

La valeur du coefficient de correlation est calculée à partir des données disponibles.

Un coefficient de correlation très élevé, mais obtenu sur peu de données est moins significatif qu'un coefficient plus faible, mais déterminée sur un grand nombre de données.

A la limite, si nous n'avions que deux observations, *R* serait égal à 1, mais aucune conclusion ne saurait en être déduite.

Obtenu sur un échantillon de taille réduite, *R* devrait être rectifié. La formule suivante est utilisée, ou *k* est le nombre de variables explicatives et *n* le nombre de données:

$$R = 1 - \frac{dispersion\,résiduelle}{dispersion\,totale}\frac{n-1}{n-k-1}$$

Le test *F* (analyse de la variance) permet d'intégrer la taille de l'échantillon dans l'appréciation de la qualité de la représentation:

$$F = \frac{\dfrac{\sum\left(\hat{Y}_i - \bar{Y}\right)^2}{k}}{\dfrac{\sum\left(\hat{Y}_i - Y_i\right)^2}{n-k-1}} = \frac{\text{dispersion expliquée moyenne}}{\text{dispersion résiduelle moyenne}}$$

Cette valeur doit être comparée à celle qui est lue dans une table de Fisher-Snédécor pour *k* degré de liberté au numérateur et *n-k-1* au dénominateur à un seuil de confiance α.

Le tableau suivant résume cette étude dite « Analyse de la variance »

Tableau d'analyse de la variance pour la régression linéaire simple (test F):

Source	Somme des carrées	Degrés de liberté	Moyenne des carrés	F
Régression	$\sum\left(\hat{Y}_i-\overline{Y}\right)^2$	k	$\frac{\sum\left(\hat{Y}_i-\overline{Y}\right)^2}{k}$	$F=\frac{\frac{\sum\left(\hat{Y}_i-\overline{Y}\right)^2}{k}}{\frac{\sum\left(\hat{Y}_i-Y_i\right)^2}{n-k-1}}$
Erreur	$\sum\left(Y_i-\hat{Y}_i\right)^2$	n-k-1	$\frac{\sum\left(Y_i-\hat{Y}_i\right)^2}{n-k-1}$	
Total	$\sum\left(Y_i-\overline{Y}\right)^2$	n-1		
F lue à partir d'une table de Fisher-Snédécor pour *k* degré de liberté au numérateur et *n-k-1* au dénominateur à un seuil de confiance α.				

VII.-2.- Validité des coefficients:

Les tests précédents permettent d'avoir une idée de la validité de la régression dans son ensemble. Il importe de connaître également la validité des coefficients du modèle, c'est-à-dire de *a* dans le cas de la régression linéaire simple.

Cette validité est vérifiée par le biais du test *t* et à travers le calcul de l'intervalle de confiance du paramètre *a*.

Si l'on admet que les valeurs à estimer à partir de différents échantillons d'observations suivent une loi de Student d'écart-type S_a, nous pouvons évaluer la probabilité que la valeur *a* soit différente de zéro. La statistique t suivante

$$t=\frac{a-0}{S_a}=\frac{a}{S_a}$$

nous donne le nombre d'écarts-type qui séparent la valeur observée de 0.

La statistique *t* mesure ainsi le degré de rareté, dans une population où la valeur de *a* est 0, d'échantillons d'observations pour lesquels $a=a_0$

L'intervalle de confiance de *a* est obtenu comme on a déjà vu. Si t_α est le nombre d'écarts-types correspondant au seuil de confiance α, il y a une probabilité $(1-\alpha)$ que la valeur de *a* soit comprise dans l'intervalle $[a-t_{\alpha/2}S_a;\ a+t_{\alpha/2}S_a]$.

VIII.- Fiabilité de la représentation:

En ajustant le nuage de points représentatifs des différents observations par une droite, nous avons admis implicitement que la relation liant *X* et *Y* était du type $Y_i=aX_i+b+e_i$ (i=1,2,…,n) où e_i est un terme d'erreur aléatoire appelé aussi le résidus et respectant les conditions suivantes:

e_i est une variable aléatoire normale de moyenne nulle,

e_i est indépendant de e_k: aucune corrélation n'existe entre les résidus,

e_i est indépendant de X_i: aucune corrélation ne peut être trouvée entre le terme d'erreur et la valeur de la variable.

Il convient de vérifier si ces conditions sont bien respectées, en particulier les deux dernières.

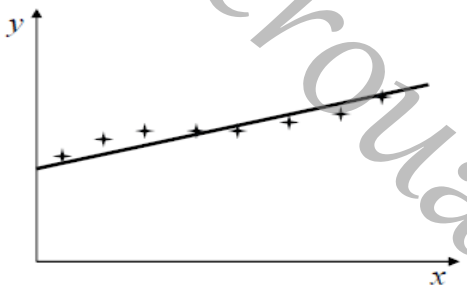
Lorsqu'il y a *autocorrélation* entre les résidus, les erreurs ne sont plus indépendantes:

L'autocorrélation positive caractérise une situation où

$e_i > 0 \Rightarrow e_{i+1} > 0$

$e_i < 0 \Rightarrow e_{i+1} < 0$.

Un tel phénomène est enregistré par exemple sur le graphique suivant:

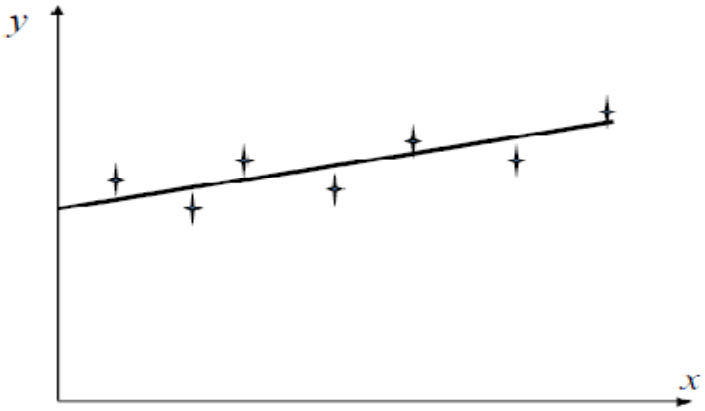


L'autocorrélation négative caractérise les situations où

$e_i > 0 \Rightarrow e_{i+1} < 0$

$e_i < 0 \Rightarrow e_{i+1} > 0$.

le graphique suivant montre un tel cas:



L'apparition d'un certain degré d'autocorrélation entre les erreurs peut indiquer que le modèle a été mal spécifié, omettant par exemple d'intégrer une variable explicative importante.

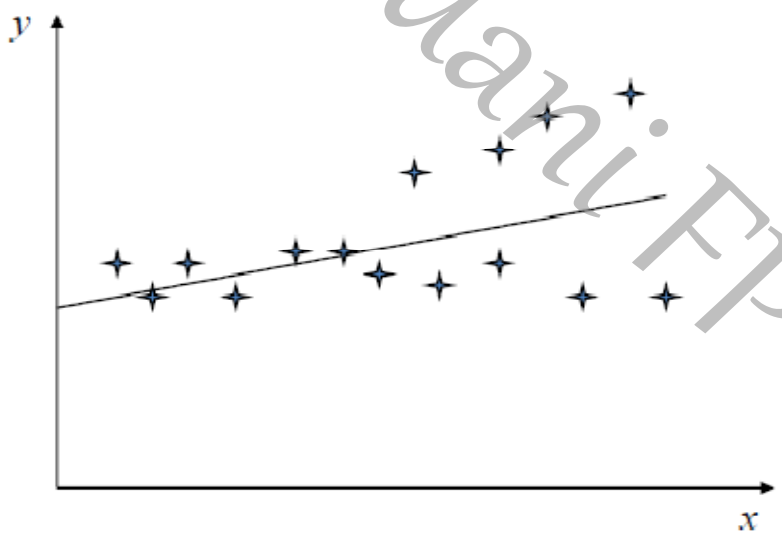
Le test de Durbin-Watson permet de repérer le degré d'autocorrélation des résidus: il demande de calculer

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

Un valeur de *d* proche de 2 manifeste l'absence d'autocorrelation, alors qu'une valeur faible correspond à une situation d'autocorrélation positive et une valeur élevée à une situation d'autocorrélation négative.

L' hétéroscédasticité est un phénomène lié au fait que la variance des erreurs n'est pas constante sur l'ensemble des observations, mais au contraire dépend de X_i.

C'est ce qui se produit par exemple si l'erreur est de plus en plus importante en valeur absolue pour des valeurs plus élevées de X_i, comme on le voit sur le graphique suivant:



L'apparaîtion de tels phénomènes altère la fiabilié des tests présentés précédement. Lorsqu'ils se produisent, la formulation du modèle doit être revue.

IX.- Prévisions à l'aide du modèle :

Un modèle de régression est construit dans le but d'expliquer à partir des observations dans quelles conditions se détermine la valeur de la variable dépendante, mais aussi de prévoir les valeurs futures de cette variable.

En fait, il faut tenir compte de ce que le modèle a été construit à partir d'un échantillon de données et qu'il existe de toute façon un certain aléa sur les relations entre *X* et *Y*.

Deux types de prévisions peuvent être posés: Prévoir la moyenne des Y pour une valeur donnée de $X=X_0$, ou prévoir une valeur individuelle de $Y=Y_0$ pour une valeur donnée de $X=X_0$.

1) Prévision par un intervalle pour la moyenne de Y en un point donné X_0 :

L'intervalle de confiance: à un seuil de confiance α, pour la moyenne de Y en $X=X_0$ est donné par:

$$\left[\hat{Y}_{X_0} - t_{1-\frac{\alpha}{2},n-2} S_e \sqrt{\frac{1}{n} + \frac{(X_0 - \overline{X})^2}{\sum X^2 - n\overline{X}^2}}; \hat{Y}_{X_0} + t_{1-\frac{\alpha}{2},n-2} Se \sqrt{\frac{1}{n} + \frac{(X_0 - \overline{X})^2}{\sum X^2 - n\overline{X}^2}}\right]$$

Où, on le rappelle, S_e est l'écart-type des erreurs du modèle, avec $t_{1-\alpha/2,n-2}$ lue dans la table de Student à n-2 degrés de liberté. L'intervalle de confiance est d'autant plus important que *Se* est élevé et *n* est faible.

2) Prévision d'un intervalle pour une observation Y_0 en un point donné X_0 :

L'intervalle de prévision ou de confiance pour une seule valeur Y0 pour une valeur donné X0 de X, est donné par :

$$\left[\hat{Y}_{X_0} - t_{1-\frac{\alpha}{2},n-2} S_e \sqrt{1 + \frac{1}{n} + \frac{(X_0 - \overline{X})^2}{\sum X^2 - n\overline{X}^2}}; \hat{Y}_{X_0} + t_{1-\frac{\alpha}{2},n-2} Se \sqrt{1 + \frac{1}{n} + \frac{(X_0 - \overline{X})^2}{\sum X^2 - n\overline{X}^2}}\right]$$

Cet intervalle de prévision est plus large que celui de la moyenne (pour une même valeur X_0 de X) parce qu'il est plus difficile en effet de prévoir une valeur individuelle que de prévoir la moyenne d'un ensemble de données.

La régression linéaire simple nous a, donc, permis de présenter les aspects principaux des techniques de régression qui peuvent être utilisées dans l'élaboration de modèles de prévision.