Matière : Méthodes statistiques
Spécialité : Chimie de l'environnement
Semestre : S2

Chapitre IV : Régression simple et multiple

L'objectif de la régression linéaire simple et multiple est d'apprendre à l'étudiant comment analyser un phénomène quelconque on utilisant des méthodes statistiques dites économétriques. En effet, la régression linéaire est une relation stochastique entre une ou plusieurs variables. Elle est appliquée dans plusieurs domaines, tels que la physique, la biologie, la chimie, l'économie…etc.

Dans un premier temps, nous allons introduire la régression linéaire où on explique une variable endogène par une seule variable exogène. A titre d'exemples, on peut citer : la relation entre la variable Prix et la variable Demande. Il s'agit de la régression linéaire régression simple. Dans un deuxième temps, nous étudierons la régression linéaire multiple qui représente la relation linéaire entre une variable endogène et plusieurs variables exogènes. Autrement dit, il s'agit de régresser linéairement une grandeur économique (variable à expliquer) sur plusieurs variables explicatives (variables exogènes). Par exemple, d'après la théorie économique, la demande d'un produit peut être expliquée par les grandeurs Prix, Revenu et Publicité.

La régression linéaire simple et multiple est un outil d'analyse qui fait appel à trois domaines scientifique, à savoir :

1. la théorie économique ;
2. l'analyse statistique ;
3. la modélisation mathématique.

I. Le modèle de régression linéaire simple

I-1 Définition du modèle de régression linéaire simple

Le modèle de régression linéaire simple est une variable endogène (dépendante) expliquée par une seule variable exogène (indépendante) mise sous forme mathématique suivante :

$$Y_t = \beta_0 + \beta_1 X_t + \varepsilon_t,\ t = 1.....n$$

avec : Yt : la variable endogène (dépendante, à expliquer) à la date t ;

Xt: la variable exogène (indépendante, explicative) à la date t ;

0 ,1 : les paramètres inconnus du modèles ;

t: l'erreur aléatoire du modèle ;

n: nombre d'observations.

- **L'ordonnée à l'origine** est la **valeur**, selon la droite déterminée, **de la variable y lorsque la variable x prend pour valeur 0**. Son estimation n'a aucun intérêt pour évaluer s'il existe une relation linéaire entre deux variables. L'ordonnée à l'origine est déterminé par la méthode, mais on ne s'en sert pas, en tout cas pas directement pour évaluer le lien entre deux variables.
- **La pente**, quant à elle, correspond à **la variation de y, lorsque x varie d'une unité**. Si, par exemple, la variable en y est exprimée en kg et que la variable en x est exprimée en cm, alors la pente correspond à la variation en kg de la variable y lorsque la variable x varie d'un cm.
- **Le signe** de la pente (<0 ou >0) nous **renseigne sur la direction de la droite**, décroissante lorsque b est <0, et croissante lorsque b est >0.
- **La valeur de la pente** renseigne sur la **vitesse d'évolution de la variable y en fonction de la variable x**, mais ne présume pas de la significativité de la relation. Il est tout à fait possible de voir une pente de valeur 10 non significative, et une pente de valeur 2 significative.

I.1. Le principe

Il consiste de **trouver la droite** (c'est-à-dire **déterminer son équation**) qui **passe au plus près de l'ensemble des points** formés par les couples (xi ; yi) figure 1.

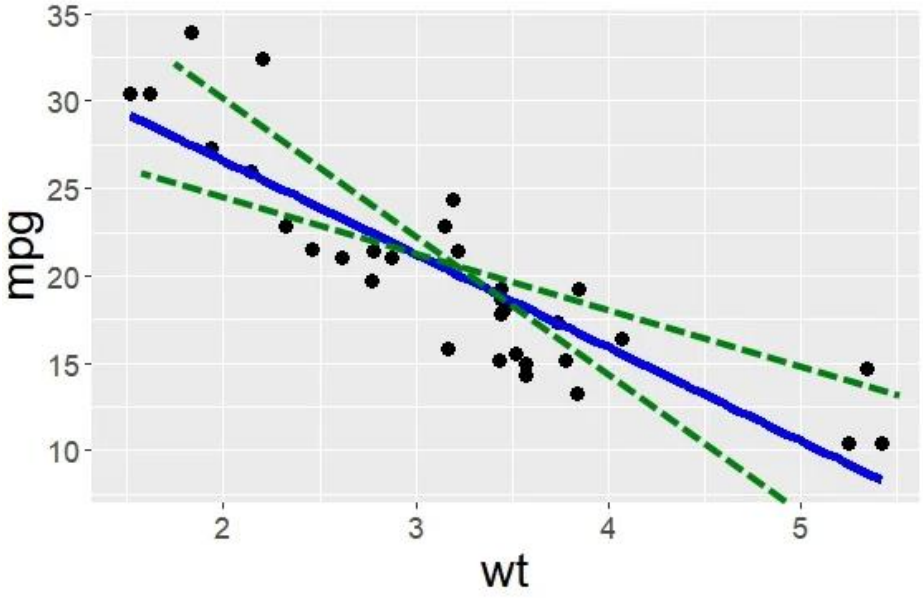


Figure 1 : Régression simple

Pour trouver cette droite qui passe au plus près de l'ensemble des points, **on mesure la distance au carré entre chaque point et chaque droite potentielle**. La droite qui passe au plus près de l'ensemble des points est celle qui **minimise la somme de ces distances au carré. C'est ce qu'on appelle la méthode des moindres carrées (ou MCO pour moindres carrés ordinaires** ou encore OLS pour Ordinary Least Squares). Ces distances entre chaque point observé et la droite déterminée par les moindres carrés sont appelées les résidus du modèle de régression figure 2.

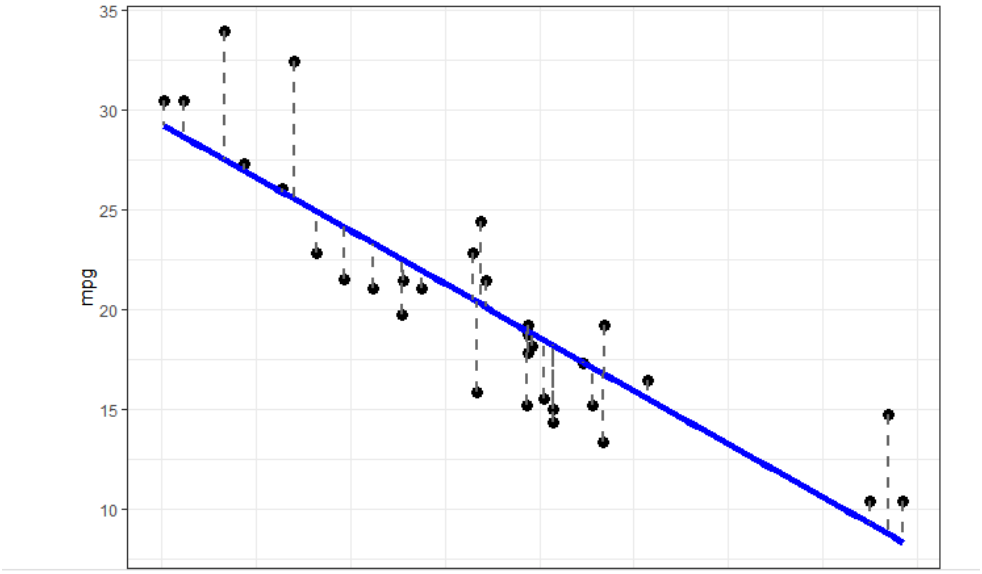


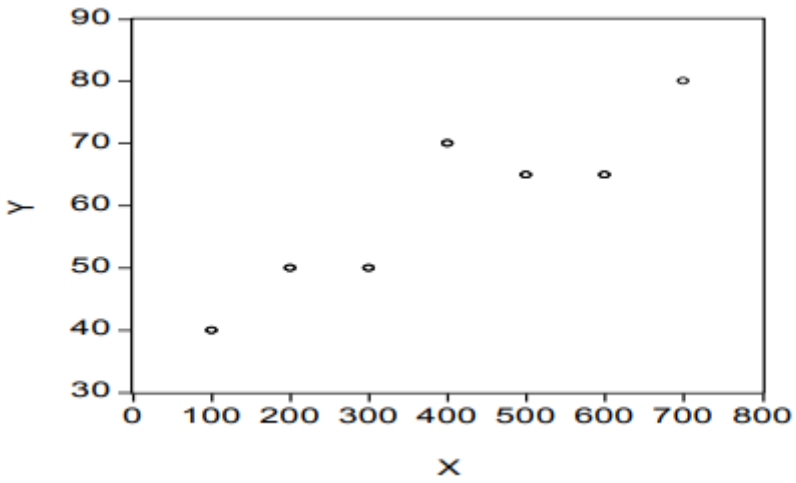
Figure 2 : les résidus du modèle de régression

- Exemple : nous disposons des données qui sont représentés dans le tableau suivant :

X_t	100	200	300	400	500	600	700
Y_t	40	50	50	70	65	65	80

Où Yt désigne les quantités consommées et Xt désigne le prix des quantités consommées.

On trace un graphique des couples de données liant le prix et les quantités Consommées. Nous



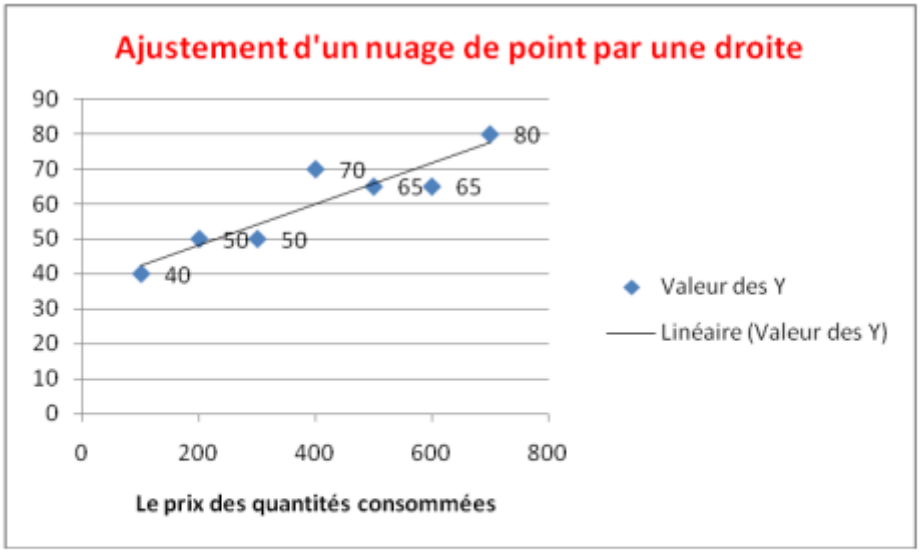
Estimation des paramètres : Nous savons que :

$$\hat{\beta}_0 = \overline{Y} - n\overline{X}$$

$$\hat{\beta}_1 = \frac{\sum_{t=1}^{n} X_t Y_t - n\overline{YX}}{\sum_{t=1}^{n} X_t^2 - n\overline{X}^2} = \frac{\sum_{t=1}^{n} (Y_t - \overline{Y})(X_t - \overline{X})}{\sum_{t=1}^{2} (X_t - \overline{X})^2}$$

Application numérique : 0 = ˆ β 36.42 et 1 ˆ β = 0.0589

Ajustement du nuage par la droite d'équation



$\hat{Y}_t = 36.42 + 0.0589 X_t$: désigne la droite qui ajuste le nuage de point.

II. Les modèles de régression multiple

La régression linéaire multiple (RLM) est une généralisation de la régression linéaire simple, dans le sens où cette approche permet de **mettre en relation une variable** réponse (variable dépendante), de type numérique continu, avec plusieurs variables explicatives (ou indépendantes), de type numérique continu, ou catégorielle par l'intermédiaire d'une fonction linéaire dans ses paramètres (addition de paramètres).

La RLM peut être utilisée à des fins de **prédiction,** mais aussi pour construire un **modèle de régression parcimonieux.** Il s'agit d'un modèle de régression simple et informatif dans lequel les coefficients obtenus seront interprétés

II- 1 : Le modèle linéaire générale

1- Présentation

Le modèle générale est une généralisation du modèle simple dans lequel figurent plusieurs variables explicatives :

$$Y_t = \beta_0 + \beta_1 x_{1t} + \beta_2 x_{2t} + \ldots\ldots\ldots\ldots + \beta_k x_{kt} + \varepsilon_t \ldots\ldots\ldots, \qquad t = 1 \ldots\ldots\ldots n$$

Avec :

Y_t = variable à expliquer a la date t

x_{1t} = variable explicative 1 à la date t .

.

.

.

x_{kt} = variable explicative k à la date t .

avec : $\beta_0, \beta_1, \beta_2 \ldots\ldots\ldots\ldots\ldots \beta_k$: paramètres du modèle.

ε_t = Erreur de spécification elle est inconnue et restera inconnue.

n = nombre d'observations

2. Estimation

L'estimation des paramètres par la méthode des moindres carrés repose sur le même principe que celui de la régression linéaire simple, mais appliqué à p dimensions. Il s'agit simplement d'une généralisation.

Ainsi, il ne s'agit plus de chercher la droite qui passe au mieux des couples des points (yi ; xi) mais le plan à p dimensions qui passe au mieux des points de coordonnées (yi, xi1, …xip).

3. Ajustement et interprétaion

La méthode des moindres carrés aboutit à une estimation ajustée des coefficients.

Ce terme **"ajustée"**, signifie **"après prise en considération des effets linéaires des autres co-variables sur la réponse,** mais aussi sur la variable prédictive".Dit autrement, le coefficient **β1**

correspond à la pente de la relation entre y et x1 lorsque les effets linéaires des autres covariables ont été enlevés, à la fois au niveau de la réponse mais aussi au niveau de x1.

Dit encore autrement, il s'agit de la pente de la relation entre l**a part du prédicteur qui est indépendante des autres prédicteurs et la réponse**, elle aussi indépendante des autres prédicteurs. Les coefficients estimés par régression linéaire multiple sont parfois nommés pentes partielles.

Le test de b = 0 équivaut à tester l'hypothèse : est ce que la réponse est associée à la variable prédictive étudiée, toutes choses étant égales par ailleurs, c'est-à-dire à niveau constant des autres variables prédictives.

4. Conditions d'application

doivent être respectées :

- **Linéarité** des relations entre la réponse et les variables prédictives (utilisation d'une transformation si besoin)
- **Normalité** des résidus
- **Homoscédasticité** des résidus
- **Absence de multicollinéarité** (forte corrélation linéaire) entre les variables prédictives. La collinéarité peut être évaluée en étudiant la corrélation de Pearson entre chaque couple de variables indépendantes.

III. Interprétation des résultats

Pour interpréter les résultats dans les deux cas, la régression linéaire simple ou multiple, on doit caractériser les points suivants :

- La valeur critique de F pour déterminer la significativité de la relation

- Le coefficient de corrélation (détermination multiple) : force du lien entre les variables indépendantes et dépendante o Le coefficient de détermination (le premier R^2 dans le tableau) : portion de la variable dépendante expliquée par les variables indépendantes
- Les coefficients: le premier est la constante de l'équation (le b0), les autres sont les coefficients (b1, … bn) des différentes variables indépendantes (X1, … Xn) :
    - Forme de l'équation de la courbe : y = b0 + b1x1 + b2x2 + … + bnxn (où n est le nombre de variables indépendantes)