République Algérienne Démocratique et Populaire

Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

UNIVERSITÉ MOHAMED KHIDER, BISKRA

FACULTÉ des SCIENCES EXACTES et des SCIENCES de la NATURE et de la VIE

DÉPARTEMENT DE MATHÉMATIQUES

Mémoire présenté en vue de l'obtention du Diplôme :

MASTER en Mathématiques

Option : Statistique

Par

BEN MAZOUZ RACHIDA

Titre :

REGRESSION POLINOMIAL ET APPLICATIONS

Membres du Comité d'Examen :

Pr.	NECIR ABDELHAKIM	UMKB	Président
Pr.	CHERFAOUI MOULOUD	UMKB	Encadreur
Dr.	CHINE AMEL	UMKB	Examinatrice

Juin 2024

Dédicace

Louange à Dieu, cela suffit, et les prières soient sur l'Élu bien-aimé, sur sa famille et sur ceux qui lui sont fidèles. Quant à ce qui suit :

Louange à Dieu qui m'a permis de valoriser cette étape de mon parcours académique avec ce mémoire, fruit de l'effort et de la réussite, grâce à lui, le Tout-Puissant, et des dons pour :

Ceux qui étaient la raison de mon existence étaient mon père, que Dieu ait pitié de lui, et ma chère mère, que Dieu la protège et prenne soin d'elle.

A mon soutien qui m'a encouragé à terminer mes études et à supporter toutes les difficultés avec moi, mon cher époux, Tilly Nabil.

Aux plaisirs de mon foie et de mes fleurs, mes chers enfants Yaacoub, Rihab et Ayoub.

A mes chers frères et sœurs.

À ceux qui m'ont soutenu alors que nous ouvrons la voie vers la réussite de notre voyage scientifique, mon compagnon sur le chemin de l'Asie, aidez

À tous ceux qui m'ont aidé et ont joué un rôle, de près ou de loin, dans la réalisation de ce travail, et en particulier, la chère Shaima chekkal et la chère Loubna Zernadji.

REMERCIEMENTS

Nous remercions Dieu Tout-Puissant, qui nous a inspiré la patience et la persévérance, nous a donné la force et la détermination pour poursuivre notre cheminement académique et nous a accordé le succès dans l'accomplissement de ce travail.

J'adresse mes sincères remerciements et mon appréciation au professeur superviseur, le **Pr. Cherfaoui Mouloud**, pour sa gentillesse dans la supervision de cette recherche et son souci de veiller à ce que ce travail soit dans une forme complète, non déformée par une quelconque lacune. Je demande à **Dieu** de le récompenser. la meilleure récompense pour tous les efforts qu'il a déployés pour nous et pour tous les conseils et orientations qu'il a gardés à l'esprit.

J'adresse également mes sincères remerciements et ma sincère gratitude au **Pr. Yahia Djebrane** pour son soutien et son aide dans cette réalisation. Je remercie également l'étudiante distingué et généreux qui a apporté une grande contribution tout au long de l'année universitaire, Zernadji Loubna.

Je ne peux également manquer d'adresser mes sincères remerciements à tous les membres du comité de discussion, chacun en son nom, et à tous mes distingués professeurs en particulier, ainsi qu'au personnel du Département de Mathématiques en général.

Table des matières

Table des figures

Liste des tableaux

Introduction

Les modèles mathématiques et statistiques sont des outils essentiels pour analyser les données et établir des relations entre les variables. Parmi ces modèles, la régression polynômiale se distingue comme l'un des modèles les plus efficaces et flexibles pour décrire les relations non linéaires entre les variables. Dans ce contexte, l'étude de la régression polynômiale revêt une importance particulière en raison de ses applications étendues dans divers domaines tels que l'économie, les sciences naturelles et la science des données.

Les racines de la régression polynômiale remontent aux travaux des mathématiciens des siècles passés. La première utilisation pratique des modèles de régression a eu lieu au $XVIIIe$ siècle, lorsque les scientifiques les ont utilisés pour décrire les phénomènes astronomiques et physiques. Avec le développement de la statistique au $XIXe$ siècle, les chercheurs ont commencé à utiliser la régression linéaire comme base pour leurs modèles, puis les études se sont étendues pour inclure la régression non linéaire, y compris la régression polynômiale.

Le XXe siècle a vu des développements importants dans ce domaine, avec le renforcement des fondements théoriques de la régression polynômiale et l'amélioration des méthodes d'estimation des paramètres et de test des modèles. Cela a coïncidé avec l'apparition des ordinateurs et des logiciels avancés qui ont facilité les calculs et les analyses, permettant ainsi l'application de la régression polynômiale à une plus

grande échelle et avec une efficacité accrue.

Nous exposons ce travail dans deux chapitres et deux annexe :

Chapitre 01 : Présente la régression linéaire simple et multiple, ainsi que certaines propriétés qui leur sont associées, et l'estimation des paramètres de chacun de ces modèles.

Chapitre 02 : Le deuxième chapitre traite la régression linéaire polynômiale et quelque propriétés fondamentales. Une deuxième partie met en pratique la régression linéaire polynômiale sur des données simulées.

Annexe : Un rappel sur le calcul de l'inverse d'une matrice a été exposé dans l'annexe A tandis que les notations utilisées au cours du mémoire sont exposées dans l'annexe B.

Chapitre 1

Régression linéaire simple et multiple

La régression est une méthode statistique qui permet d'établir la relation entre une variable quantitative a expliquée (endogène ou dépendante) à une ou plusieurs autres variables quantitatives explicatives (exogènes ou indépendantes), sous la forme d'un modèle. Elle est dite linéaire si elle impose une forme fonctionnelle linéaire dans les paramètres du modèle. Si on s'intéresse à la relation entre deux variables on parlera de la régression linéaire simple et si la relation porte entre une variable et plusieurs autres variables, on parlera alors de la régression linéaire multiple. Cette dernière sert à prévoir les valeurs futures de l'une des variables en fonction de l'autre.

Nous introduisons dans ce chapitre la notion de régression d'une variable réelle Y par rapport à une variable X par une droite (régression linéaire simple). Ensuite, nous présentons une généralisation de la notion de droite, en remplaçant X par plusieurs variables $X_1, X_2, ..., X_p$ (régression linéaire multiple), chargées de permettre la prévision linéaire de Y.

1.1 La régression linéaire simple

La régression linéaire simple est une méthode statistique permettant de trouver une relation linéaire entre une variable explicative X et une variable à expliquer Y. Ce modèle consiste à considérer Y comme une fonction affine de X. En d'autre terme, la régression linéaire a pour but de trouver une droite ajustée au nuage de points de Y en fonction de X.

Définition 1.1.1 (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 la forme mathématique suivante :*

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \ \ i = 1, ..., n,$$

avec,

y_i :	la $i^{éme}$ observation de la variable aléatoire à expliquer$Y$,
x_i :	la $i^{éme}$ observation de la variable explicative X,
β_0 et β_1 :	sont des constantes inconnues appelées paramètres du modèle,
ε_i :	l'erreur (ou bruit) aléatoire du modèle,
n :	la taille de l'échantillon.

Hypothèses du modèle : En générale, le modèle repose sur les hypothèses suivantes :

1. $E(\varepsilon_i) = 0$, l'erreur centrée.
2. $E(\varepsilon_i^2) = \sigma_\varepsilon^2 < \infty$, $i = 1, ..., n$, la variance de l'erreur est constante (l'hypothèse d'homoscédasticité).
3. $cov(\varepsilon_i, \varepsilon_{i'}) = 0$, *si* $\varepsilon_i \neq \varepsilon_{i'}$, les erreurs ne sont pas auto-corrélées .
4. $cov(x_i, \varepsilon_{i'}) = 0$, l'erreur n'est pas corrélée avec la variable exogène.

5. La variable exogène n'ai pas aléatoire.

6. Le modèle est linéaire en X par rapport aux paramètres.

1.1.1 Estimation des paramètres par la méthode MCO

Soit le modèle suivant

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

L'estimation des paramètres β_0, β_1 sont $\hat{\beta}_0,\ \hat{\beta}_1$ définissant la droite de régression

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1\ x_i.$$

Les valeurs $\hat{\beta}_0$, $\hat{\beta}_1$ sont obtenues en minimisant la somme des carrés des erreurs :

$$\min S^2 = \min \sum_{i=1}^{n} \varepsilon_i^2 = \min \sum_{i=1}^{n} \left(Y_i - \beta_0 - \beta_1 X_i\right)^2.$$

Pour que cette fonction ait un minimum, il faut que les dérivées par rapport à β_0 et β_1 soient nuls, alors

$$\frac{\partial S}{\partial \beta_0} = 0 \Leftrightarrow 2\sum_{i=1}^{n} \left(Y_i\ - \beta_0 - \beta_1 X_i\right)(-1) = 0 \Longrightarrow \sum_{i=1}^{n} Y_i = n\beta_0 + \beta_1 \sum_{i=1}^{n} X_i, \tag{1.1}$$

$$\frac{\partial S}{\partial \beta_1} = 0 \Leftrightarrow 2\sum_{i=1}^{n} \left(Y_i\ - \beta_0 - \beta_1 X_i\right)(-X_i) = 0 \Longrightarrow \sum_{i=1}^{n} Y_i X_i = \beta_0 \sum_{i=1}^{n} X_i + \beta_1 \sum_{i=1}^{n} X_i^2. \tag{1.2}$$

Les quantités $\hat{\beta}_0$ et $\hat{\beta}_1$ sont les solutions des équations (1.1) et (1.2). Ainsi, on obtient d'après (1.1) :

$$\hat{\beta}_0 = \frac{1}{n}\sum_{i=1}^{n} Y_i - \frac{1}{n}\hat{\beta}_1 \sum_{i=1}^{n} X_i,$$

ou bien

$$\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{X}, \text{ avec } \overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i \text{ et } \overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

En remplaçant la valeur de $\hat{\beta}_0$ dans l'équation (1.2), on obtient :

$$\sum_{i=1}^{n} Y_i X_i - \overline{Y}\sum_{i=1}^{n} X_i = \hat{\beta}_1 \left(\sum_{i=1}^{n} X_i^2 - \overline{X}\sum_{i=1}^{n} X_i \right),$$

d'où

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} Y_i X_i - \overline{Y}\sum_{i=1}^{n} X_i}{\sum_{i=1}^{n} X_i^2 - \overline{X}\sum_{i=1}^{n} X_i} = \frac{\sum_{i=1}^{n} Y_i X_i - n\overline{XY}}{\sum_{i=1}^{n} X_i^2 - n\overline{X}^2} = \frac{\sum_{i=1}^{n} (Y_i - \overline{Y})(X_i - \overline{X})}{\sum_{i=1}^{n} (X_i - \overline{X})^2},$$

ou encore :

$$\hat{\beta}_1 = \frac{Cov\,(X,Y)}{Var\,(X)}, \quad \text{et} \quad \hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{X},$$

avec,

$$Cov\,(X,Y) = \frac{1}{n}\sum_{i=1}^{n} (X_i - \overline{X})(Y_i - \overline{Y}) = \left[\frac{1}{n}\sum_{i=1}^{n} X_i Y_i \right] - \overline{XY}.$$

1.1.2 Différentes écritures du modèle

Le modèle théorique (modèle non ajusté) Le modèle théorique est définie par :

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

Le modèle estimé (modèle ajusté) Le modèle estimé est définie par :

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i.$$

avec,

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i, \text{ et } \varepsilon_i = Y_i - \hat{Y}_i = Y_i - \hat{\beta}_0 + \hat{\beta}_1 X_i,$$

ε_i : est le résidu du modèle.

Exemple 1.1.1 *L'analyse de la température de fonctionnement d'un procédé chimique sur le rendement d'un produit a donné les valeurs suivantes pour la température X_i et le rendement correspondant Y_i :*

Température C° (X_i)	100	200	300	400	500	600	700
Rendement (Y_i)	40	50	50	70	65	65	80

On trace un graphique des couples de données liant la température et le rendement. Le graphe ci-dessous représente les points (X_i, Y_i) pour ces données et suggère une relation linéaire entre X et Y.

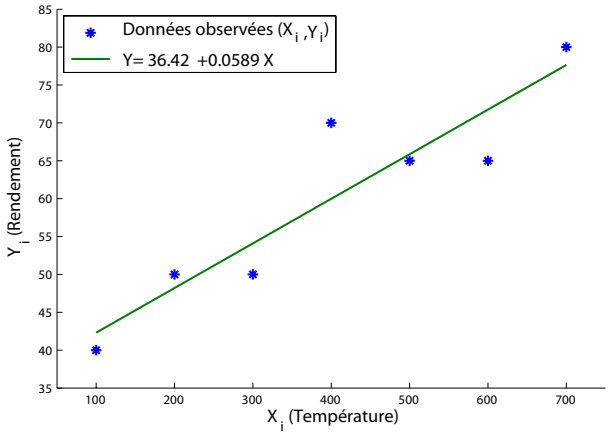


FIGURE 1.1 – Nuage des points des données observées

- **Estimation des paramètres :** Nous savons que

$$\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{X}.$$

$$\hat{\beta}_1 = \frac{\frac{1}{n}\sum_{i=1}^{n} Y_i X_i - \overline{XY}}{\frac{1}{n}\sum_{i=1}^{n} X_i^2 - \overline{X}^2}.$$

- **Application numérique :**

$$\hat{\beta}_0 = 36.42, \text{ et } \hat{\beta}_1 = 0.0589.$$

- Ajustement du nuage par la droite d'équation

$$\hat{Y}_i = 36.42 + 0.0589X_i.$$

Cette droite désigne la droite qui ajuste le nuage de point.

1.1.3 Estimation de la variance des erreurs σ^2

Soit le modèle de régression linéaire simple :

$$Y_i = \beta_0 + \beta_1 X + \varepsilon_i.$$

Sachant que, pour tout $i = \overline{1,n}$ les résidus sont :

$$\begin{aligned} e_i &= Y_i - \hat{Y}_i, \\ \hat{Y}_i &= \hat{\beta}_0 + \hat{\beta}_1 X_i, \\ \overline{Y} &= \beta_0 + \beta_1 \overline{X} + \overline{\varepsilon_i} \end{aligned}$$

Alors,

$$e_i = (\beta_0 + \beta_1 X + \varepsilon_i) - \left(\hat{\beta}_0 + \hat{\beta}_1 X_i\right),$$

on remplace $\hat{\beta}_0$ par sa valeur, on obtient :

$$e_i = \beta_0 + \beta_1 X_i - \overline{Y} + \hat{\beta}_1 \overline{X} - \hat{\beta}_1 X_i + \varepsilon_i,$$

On remplace aussi $\overline{Y}$ par sa valeur on obtient :

$$\begin{aligned} e_i &= \left(\beta_0 - \hat{\beta}_1\right) X_i - \left(\beta_0 - \hat{\beta}_1\right) \overline{X} + \varepsilon - \overline{\varepsilon} = \left(\beta_0 - \hat{\beta}_1\right) \left(X_i - \overline{X}\right) + \varepsilon - \overline{\varepsilon}, \\ &= \left(\beta_0 - \hat{\beta}_1\right) x_i + \varepsilon - \overline{\varepsilon}, \text{ car } x_i = \left(X_i - \overline{X}\right). \end{aligned}$$

D'où

$$\begin{aligned} \sum_{i=1}^{n} e_i^2 &= \sum_{i=1}^{n} \left[\left(\beta_1 - \hat{\beta}_1\right) x_i + (\varepsilon_i - \overline{\varepsilon})\right]^2, \\ &= \sum_{i=1}^{n} (\varepsilon_i - \overline{\varepsilon})^2 + \left(\beta_1 - \hat{\beta}_1\right)^2 \sum_{i=1}^{n} x_i^2 + 2\left(\beta_1 - \hat{\beta}_1\right) \sum_{i=1}^{n} x_i (\varepsilon_i - \overline{\varepsilon}). \end{aligned}$$

On passant à l'espérance $E\left(\sum_{i=1}^{n} e_i^2\right)$, on obtient :

$$E\left(\sum_{i=1}^{n} e_i^2\right) = (n-1)\,\sigma_\varepsilon^2 + \sigma_\varepsilon^2 - 2\sigma_\varepsilon^2 = (n-2)\,\sigma_\varepsilon^2,$$

on déduit :

$$\sigma_\varepsilon^2 = \frac{E\left(\sum_{i=1}^{n} e_i^2\right)}{n-2}.$$

Finalement :

$$\hat{\sigma}_\varepsilon^2 = \frac{\sum_{i=1}^{n} e_i^2}{n-2},$$

tel que, $\hat{\sigma}_\varepsilon^2$ est un estimateur sans biais.

1.1.4 Propriétés statistiques des estimateurs

Propriétés 1.1.1 *1. $\hat{\beta}_1, \hat{\beta}_0$ sont des estimateurs sans biais ($E\left(\hat{\beta}_1\right) = \beta_1, E\left(\hat{\beta}_0\right) = \beta_0$).*

2. $var\left(\hat{\beta}_1\right) = \frac{\sigma_\varepsilon^2 \sum\limits_{i=1}^{n} x_i^2}{\left(\sum\limits_{i=1}^{n} x_i^2\right)^2} = \frac{\sigma_\varepsilon^2}{\sum\limits_{i=1}^{n} x_i^2}$.

3. $var\left(\hat{\beta}_0\right) = \sigma_\varepsilon^2 \left[\frac{\sum\limits_{i=1}^{n} X_i^2}{n\sum\limits_{i=1}^{n} x_i^2}\right] = \sigma_\varepsilon^2 \left[\frac{\sum\limits_{i=1}^{n} X_i^2}{n\sum\limits_{i=1}^{n}\left(X_i - \overline{X}\right)^2}\right]$.

4. $cov\left(\hat{\beta}_0, \hat{\beta}_1\right) = -\frac{\overline{X}\sigma_\varepsilon^2}{\sum\limits_{i=1}^{n} x_i^2}$.

Test des coefficients et les intervalles de confiances

Test des coefficients

Dans notre cas on sait que

$$\hat{\sigma}_\varepsilon^2 = \frac{\sum\limits_{i=1}^{n} e_i^2}{n-2} \text{ et } \varepsilon_i \to \mathcal{N}\left(0, \sigma_\varepsilon^2\right). \tag{1.3}$$

Alors,

$$(n-2)\,\hat{\sigma}_\varepsilon^2 = \sum_{i=1}^{n} e_i^2 \Longrightarrow \frac{(n-2)\,\hat{\sigma}_\varepsilon^2}{\sigma_\varepsilon^2} = \frac{\sum\limits_{i=1}^{n} e_i^2}{\sigma_\varepsilon^2} \to \chi^2_{n-2}$$

D'un autre coté on a :

$$\hat{\beta}_1 \to \mathcal{N}\left(\beta_1, \frac{\sigma_\varepsilon^2}{\sum\limits_{i=1}^{n} x_i^2}\right) \text{ et } \hat{\beta}_0 \to \mathcal{N}\left(\beta_0, \sigma_\varepsilon^2\left(\frac{\sum\limits_{i=1}^{n} x_i^2}{n\sum\limits_{i=1}^{n}\left(X_i - \overline{X}\right)^2}\right)\right)$$

D'où on obtient les variables centrées réduites Z_1 et Z_2 :

$$Z_1 = \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\frac{\sigma_\varepsilon^2}{\sum\limits_{i=1}^{n} x_i^2}}} \to \mathcal{N}(0,1) \text{ et } Z_2 = \frac{\hat{\beta}_0 - \beta_0}{\sqrt{\sigma_\varepsilon^2\left(\frac{\sum\limits_{i=1}^{n} x_i^2}{n\sum\limits_{i=1}^{n}\left(X_i - \overline{X}\right)^2}\right)}} \to \mathcal{N}(0,1).$$

En se basant sur la définition de la loi de Student qui est : le rapport d'une loi normale

centrée réduite et la racine carrée d'une loi de Khi-deux divisée par son degrés de liberté on peut justifier les deux résultats suivants :

$$T_{\hat{\beta}_1} = \frac{\hat{\beta}_1 - \beta_1}{\hat{\sigma}_{\hat{\beta}_1}} \to T_{n-2}. \tag{1.4}$$

et

$$T_{\hat{\beta}_0} = \frac{\hat{\beta}_0 - \beta_0}{\hat{\sigma}_{\hat{\beta}_0}} \to T_{n-2}. \tag{1.5}$$

À partir des résultats (1.4) et (1.5), on peut effectuer les tests d'hypothèses suivants :

$$\begin{cases} H_0 : \beta_0 = 0 \\ H_1 : \beta_0 \neq 0 \end{cases} \quad \text{et} \quad \begin{cases} H_0 : \beta_1 = 0 \\ H_1 : \beta_1 \neq 0 \end{cases}$$

Test sur le paramètre β_1

Le test du paramètre β_1 consiste à tester l'hypothèse suivante :

$$\begin{cases} H_0 : \beta_1 = 0 \\ H_1 : \beta_1 \neq 0 \end{cases}$$

Ce test appelé test bilatéral.

Sous l'hypothèse $H_0 : \beta_1 = 0$, on obtient la valeur critique (T_C) tel que :

$$T_C = \left| \frac{\hat{\beta}_1}{\hat{\sigma}_{\hat{\beta}_1}} \right| \to T_t \left(n - 2, \alpha/2 \right),$$

avec :

- T_C : désigne la valeur de la réalisation de la statistique $\left(T_{\hat{\beta}_1} \right)$ (dite calculée).
- $\hat{\beta}_1$: désigne la valeur estimée du paramètreβ_1.
- $\hat{\sigma}_{\hat{\beta}_1}$: désigne la valeur de l'écart–type du paramètre β_1.

- α : le seuil de decision donné.
- $n-2$: le degré de liberté de la loi de Student.
- $T_t\,(n-2,\alpha/2)$: désigne la valeur critique du test qui correspond au quantile d'ordre $\alpha/2$ d'une loi de Student à $n-2$ ddl (T_t peut être lue, par exemple, à partir de la table statistique de la loi de Student).

Règle de décision

• Si $|T_C| < T_t\,(n-2,\alpha/2)$ on ne rejette pas l'hypothèse H_0 : dans ce cas on conclut que la variable x_i n'est pas contributive à l'explication de Y.

• Si $|T_C| > T_t\,(n-2,\alpha/2)$ on rejette l'hypothèse H_0 : dans ce cas on conclut que la variable x_i est contributive à l'explication de Y.

Remarque 1.1.1 *De la même manière pour le test du paramètre β_0.*

Intervalles de confiances des paramètres β_1 et β_0

Les intervalles de confiances des paramètres β_1 et β_0 au seuil donné α (au niveau de confiance $(1-\alpha)$) sont donnés, respectivement, par :

$$P\left[\hat{\beta}_1 - T_{\alpha/2} \times \hat{\sigma}_{\hat{\beta}_1} < \beta_1 < \hat{\beta}_1 + T_{\alpha/2} \times \hat{\sigma}_{\hat{\beta}_1}\right] = 1-\alpha$$

$$P\left[\hat{\beta}_0 - T_{\alpha/2} \times \hat{\sigma}_{\hat{\beta}_0} < \alpha < \hat{\beta}_0 + T_{\alpha/2} \times \hat{\sigma}_{\hat{\beta}_0}\right] = 1-\alpha$$

avec $T_{\alpha/2}$ désigne le quantile d'ordre $\alpha/2$ d'une loi de Student à $n-2$ ddl.

1.1.5 Qualité d'Ajustement

Pour juger la qualité d'ajustement du modèle nous utilisons l'équation de l'analyse de la variance c.-à-d. cherchons tout d'abord à décomposer la variance des y_i autour de leur moyenne en une somme de deux autre variances.

$$\underbrace{\sum_{i=1}^{n}(y_i-\bar{y})^2}_{SCT}=\underbrace{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}_{SCR}+\underbrace{\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2}_{SCE}, \tag{1.6}$$

où

$SCT=\sum_{i=1}^{n}(y_i-\bar{y})^2$ **:** somme des carrés totale (ou variation totale des y_i)

$SCR=\sum_{i=1}^{n}(y_i-\hat{y}_i)^2=\sum_{i=1}^{n}e_i^2$ **:** somme des carrés résiduelle (ou variation des résidus $\hat{\varepsilon}_i$ dite aussi variation résiduelle)

$SCE=\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2$ **:** somme des carrés expliqués (variation expliquée).

Coefficient de corrélation et coefficient de détermination

En probabilité et en statistique, étudier la corrélation entre deux ou plusieurs variables aléatoire, c'est étudier l'intensité de la liaison qui peut être existée entre ces variable. Une mesure de cette corrélation dans le cadre linéaire est obtenue par le calcul du coefficient appelé coefficient de corrélation. Ce coefficient est égal au rapport de leurs covariances et du produit non nul de leurs écarts types :

$$\rho=Cor\,(x,y)=\frac{Cov\,(x,y)}{\sigma_x\sigma_y}=r\,(x,y)\,.$$

avec,

$$\sigma_x^2 = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \overline{X}\right)^2,$$
$$\sigma_y^2 = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \overline{Y}\right)^2,$$

et

$$Cov\,(x,y) = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \overline{X}\right)\left(y_i - \overline{Y}\right) = \left[\frac{1}{n}\sum_{i=1}^{n} x_i y_i\right] - \overline{XY}.$$

Le coefficient de corrélation est toujours compris entre -1 et $+1$. De plus, son signe donne le sens de la corrélation où le signe positif indique que les deux variables sont proportionnelles dans le même sens, tandis que le signe négatif indique que les deux variables sont inversement proportionnelles.

Plus $|\rho|$ est près de 1, plus la corrélation est grande donc le modèle linéaire décrit bien le phénomène étudié. Par contre, si $|\rho|$ est près de zéro le modèle linéaire n'est pas adéquat pour la modélisation du problème étudié.

Pour mieux juger la qualité d'une régression linéaire, on définit un autre indicateur compris entre 0 et 1, nommé : **Coefficient de détermination**, noté R^2 est définis par :

$$R^2 = \rho^2.$$

Par conséquent, $0 \leq R^2 \leq 1$,. Ainsi, plus la valeur de R^2 est proche de 1, plus le modèle est plus significatif.

De l'équation (1.6) on peut déduire une autre expression pour le coefficient de détermination.

$$(1.6) \Longrightarrow \frac{SCT}{SCT} = \frac{SCE}{SCT} + \frac{SCR}{SCT} \Longrightarrow 1 = \frac{SCE}{SCT} + \frac{SCR}{SCT},$$

d'où

$$R^2 = 1 - \frac{SCR}{SCT} = \frac{SCR}{SCT}.$$

Analyse de la variance

L'information donnée par les valeurs SCT, SCE et SCR est présentée dans un tableau d'analyse de la variance à un seul facteur :

Source de variation	Somme des carrés	d.d.l	Moyenne des carrés	F
Expliquée	SCE	1	$MCE = \frac{SCR}{1}$	$F = \frac{MCE}{MCR}$
Résidus	SCR	$n-2$	$MCR = \frac{SCR}{n-2}$	
Total	SCT	$n-1$		

• **Test de Fisher**

Nous acceptons l'hypothèse de signification globale du modèle si :

$$F = \frac{SCE/1}{SCR/\left(n-2\right)} > f_{(1,n-2,1-\alpha)},$$

avec : $f_{(1,n-2,1-\alpha)}$ est le fractile d'ordre $1-\alpha$ de la loi de Fisher $F\left(1, n-2\right)$ lue, par exemple, à partir de la table statistique de Fisher.

1.2 Modèle de régression linéaire multiple

Le modèle de régression linéaire multiple constitue une généralisation de la régression linéaire simple, et qui sert à la mise en oeuvre de l'étude des données multidimensionnelle (plus de dimension 2).

Le modèle général représente une extension du modèle simple, avec plusieurs variables

explicatives et qui se présente sous la forme :

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + ... + \beta_j x_{j,i} + ... + \beta_p x_{p,i} + \varepsilon, \ i = 1, ..., n, \\ &= \beta_0 + \sum_{j=1}^{p} \beta_j x_{j,i} + \varepsilon_i, \end{aligned}$$

avec :

Y_i : variable à expliquer à la date i.

x_{1i} : variable explicative 1 à la date i .

$x_{j,i}$: sont des variables déterministes,

les paramètres du modèle β_j $\left(j = \overline{0, p}\right)$ sont des constantes inconnus,

ε_i : sont des termes d'erreur d'une variable aléatoire $\varepsilon = (\varepsilon_1, \varepsilon_2, ..., \varepsilon_n)^t$.

n : nombre d'observations

Le modèle de régression linéaire multiple prend la forme matricielle suivante :

$$Y = X\beta + \varepsilon,$$

avec :

Y : est un vecteur aléatoire de dimension n,

X : est une matrice de dimension $(n; p + 1)$ c'est-à-dire $X \in \mathbb{R}^n \times \mathbb{R}^{p+1}$, contient l'ensemble des observations sur les variables explicatives et la première colonne formée par la valeur 1 indique la constante β_0 dans l'équation, $\beta = (\beta_0, \beta_1, ..., \beta_p)^t$ est le vecteur de dimension $p + 1$ des paramètres du modèle et " est le vecteur des erreurs de dimension n. D'où

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{1,1} & x_{1,2} & x_{1,p} \\ 1 & x_{2,1} & x_{2,2} & x_{2,p} \\ & & & \\ 1 & x_{n,1} & x_{n,2} & x_{n,p} \end{bmatrix} \times \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

$$Y = X \times \beta + \varepsilon$$

Les hypothèses concernant le modèle sont :

(**H1**) La matrice X est de plein rang, c'est-à-dire

$$rang(X) = p + 1.$$

(**H2**) Les erreurs sont centrées, de même variance, et non corrélées entre elles

$E[\varepsilon] = 0_n,\ E(\varepsilon_i^2) = \sigma_\varepsilon^2$ (pour tout i la variance de l'erreur est constante), $Var(\varepsilon) = \sigma_\varepsilon^2 I_n,$

avec : I_n la matrice identité d'ordre n.

(**H3**) $E(\varepsilon_i \varepsilon_{i\prime}) = 0$ *si* $i \neq i\prime$ (indépendance des erreurs).

(**H4**) Les erreurs sont indépendantes des X_j $\left(j = \overline{1,p}\right)$.

1.2.1 Estimation et propriétés des estimateurs

À partir de la connaissance des valeurs X_j, on estime les paramètres inconnues du modèle (le vecteur des paramètre β et la variance σ_ε^2) par minimisation des moindres carrés.

Estimation des coefficients de régression

Construction des estimateurs par la méthode des moindres carrés. En suivant les mêmes étapes que l'estimation des paramètres pour régression linéaire simple, nous

trouvons :

$$\hat{Y} = X\hat{\beta} + \varepsilon, \text{ tel que } \hat{\beta} = \begin{pmatrix} \hat{\beta}_0 \\ \vdots \\ \hat{\beta}_i \\ \vdots \\ \hat{\beta}_p \end{pmatrix} \text{ et } \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_i \\ \vdots \\ \varepsilon_n \end{pmatrix}.$$

Trouver $\hat{\beta}$ tel que $\sum \varepsilon_i^2$ soit minimale, c'est-à-dire

$$\hat{\beta} = \arg\min_{\beta} \sum_{i=1}^{n} \varepsilon_i^2.$$

Pour calculer $\hat{\beta}$, nous proposons l'écriture matricielle suivante :

$$\begin{aligned} S(\beta) = \varepsilon'\varepsilon = ||Y - X\beta||^2 = (Y - X\beta)'(Y - X\beta) = (Y' - \beta'X')(Y - X\beta), \\ = Y'Y - Y'X\beta - \beta'X'Y + \beta'X'X\beta, \end{aligned}$$

et puisque $Y'X\beta = \beta'X'Y \in \mathbb{R}$, on obtient donc

$$S(\beta) = Y'Y - 2\beta'X'Y + \beta'X'X\beta.$$

En dérivant $S(\beta)$ par rapport à chacun des paramètres $\beta_0, \beta_1, ...\beta_p$, on obtient le système des équations :

$$\begin{aligned} -2X'Y + 2X'X\hat{\beta} = 0 \Rightarrow X'Y + X'X\hat{\beta} = 0 \Rightarrow X'Y = X'X\hat{\beta}, \\ \Rightarrow (X'X)^{-1}X'Y = (X'X)^{-1}(X'X)\hat{\beta}, \end{aligned}$$

d'où on tire finalement :

$$\hat{\beta} = (X'X)^{-1}X'Y.$$

Proposition 1.2.1 *$\hat{\beta}$ est un estimateur sans biais de β, de matrice de variance-*

covariance

$$var(\hat{\beta}) = \sigma_\varepsilon (X'X)^{-1}.$$

Remarque 1.2.1

1. *La matrice $X'X$ est carrée d'ordre $p+1$, symétrique et inversible car X est de rang $p+1$.*
2. *$X'X$ est définie positive.*
3. *$\hat{Y} = X\hat{\beta}$ est la valeur ajusté de Y.*
4. *Le vecteur des résidus $\hat{\varepsilon}$, tel que*

$$\hat{\varepsilon} = Y - \hat{Y} = Y - X\hat{\beta} = Y - X(X'X)^{-1}X'Y = (I_n - X(X'X)^{-1}X')Y.$$

Si on pose $H = X(X'X)^{-1}X'$ et $P = I_n - H$ alors

$$\hat{\varepsilon} = (I_n - H)Y = PY.$$

Proposition 1.2.2 *Les matrices H et P sont symétriques et du rang $(p+1)$ et $(n-p-1)$ respectivement, on dit alors qu'elles sont idempotentes.*

Théorème 1.2.1 *Les vecteurs $\hat{\beta}$ et $\hat{\varepsilon}$ ne sont pas corrélés entre eux.*

Théorème 1.2.2 *Le vecteur des résidus $\hat{\varepsilon}$ vérifié :*

1. *$E(\hat{\varepsilon}) = 0$.*
2. *$E(\varepsilon - \hat{\varepsilon}) = 0$.*
3. *$var(\hat{\varepsilon}) = \sigma^2 P$.*

1.2.2 Qualité d'ajustement

De même que pour la régression linéaire simple, on a une égalité triangulaire généralisée

$$\begin{array}{ccccc} \sum_{i=1}^{n}(y_i-\overline{y})^2 & = & \sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2 & + & \sum_{i=1}^{n}(y_i-\hat{y})^2, \\ SCT & = & SCE & + & SCR. \end{array}$$

La qualité d'ajustement est jugé par le coefficient de détermination R^2, défini par :

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2} = 1 - \frac{\sum_{i=1}^{n}\hat{\varepsilon}_i^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2} = 1 - \frac{\hat{\varepsilon}'\hat{\varepsilon}}{\sum_{i=1}^{n}\hat{\varepsilon}_i^2}.$$

Il est possible de prendre de plus en considération le coefficient de détermination ajusté $\overline{R^2}$, défini par :

$$\overline{R^2} = 1 - \frac{n-1}{n-p-1}(1-R^2).$$

1.2.3 Lois des Estimateurs

– On ajoutera l'hypothèse de normalité des erreurs

$$\varepsilon \to \mathcal{N}_n(0, \sigma_\varepsilon^2 I_n).$$

– Y est donc de loi normale dans $\mathbb{R}^n$

$$Y \to \mathcal{N}_n(X\beta, \sigma_\varepsilon^2 I_n).$$

– Loi de $\hat{\beta}$ est

$$\hat{\beta} \to \mathcal{N}_{p+1}(\beta, \sigma_\varepsilon^2 (X'X)^{-1}).$$

– Loi de $(n-p-1)\frac{\hat{\sigma}_\varepsilon^2}{\sigma_\varepsilon^2}$:

$$(n-p-1)\frac{\hat{\sigma}_\varepsilon^2}{\sigma_\varepsilon^2} \to \mathcal{X}^2_{n-p-1}.$$

1.2.4 Intervalles de Confiances

Notons V_j le $j^{ème}$ terme diagonal de la matrice $(X'X)$, si σ_ε^2 est connue alors

$$\frac{\hat{b}_j - b_j}{\sigma_\varepsilon \sqrt{V_j}} \to \mathcal{N}(0,1)$$

Si σ_ε^2 est inconnue, alors on déduit ce qui suit

$$\frac{\hat{b}_j - b_j}{\sigma_\varepsilon \sqrt{V_j}} \to \mathcal{T}_{n-p-1},$$

avec, $\mathcal{T}_{n-p-1}$ la loi de Student à $(n-p-1)$ degrés de liberté. Ce qui permet d'obtenir une intervalle de confiance pour tout b_j et d'effectuer des tests d'hypothèses.

Puisque en pratique rarement que σ_ε^2 est connue alors $\frac{\hat{b}_j - b_j}{\sigma_\varepsilon \sqrt{V_j}} \to \mathcal{T}_{n-p-1}$, par conséquent l'intervalle de confiance au niveau de confiance $1-\alpha$ pour chaque coefficient b_j du modèle, est donnée comme suit :

$$b_j = \left[\hat{b}_j \pm t\sigma_\varepsilon \sqrt{V_j}\right],$$

avec, t étant le fractile d'ordre $(1-\alpha/2)$ de la loi de Student à $(n-p-1)$ ddl.

Chapitre 2

Régression linéaire polynômiale

2.1 Régression polynômiale

La régression polynômiale est une méthode statistique utilisée pour représenter une relation non linéaire entre la réponse (y) et la ou les variables explicatives (x) de forme non linéaire.

Ces modèles de régression intègrent des polynômes dans leurs équations pour prendre en compte cette forme non-linéaire de la relation entre y et x :

Malgré leur capacité à modéliser des relations de formes non-linéaires (courbure, sinusoïde, etc.), les modèles polynomiaux font partie de la famille des modèles linéaires. Étant donné que le mot "modèle linéaire" désigne les paramètres du modèle et le fait que leurs effets sont multipliés.

Les modèles polynomiaux les plus couramment employés en pratique sont ceux de degré 2 (quadratique), et plus rarement de degré 3 (cubique), c'est-à-dire de la forme :

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \varepsilon,$$

et

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \varepsilon.$$

Remarque 2.1.1 *La régression linéaire est une régression polynômiale de degré* 1.

Définition 2.1.1 *On parle de régression polynômiale quand f est définie comme un polynôme par rapport à X. Cette représentation polynômiale peut être motivée par l'intuition, le contexte de l'expérience d'où émanent les données ou des critères statistiques précis, et on écrit*

$$Y = f(X, \beta_0, \beta_1, ..., \beta_p) + \varepsilon = \beta_0 + \beta_1 X + \beta_2 X^2 + ... + \beta_p X^p + \varepsilon. \qquad (2.1)$$

2.2 Principes de la régression polynômiale

La régression polynômiale est une extension de la régression linéaire multivariée. Elle permet de lier les variables par un polynôme de degré k. Un polynôme est tout simplement la somme de plusieurs expressions de la forme ax^k, où a est un nombre réel (ou complexe) et k un entier naturel. On ne dirait pas comme ça, mais un polynôme est un objet mathématique drôlement pratique puisqu'il permet d'introduire de la non-linéarité dans les relations entre variables.

En pratique, les calculs sont les mêmes que dans le cas de la régression multiple traditionnelle (la régression linéaire n'est en fait qu'une régression polynômiale de degré). L'introduction de termes polynômiaux dans un modèle de régression permet donc de modéliser simplement des relations potentiellement très complexes.

Remarque 2.2.1 *Attention, même si elle peut formidablement améliorer l'ajustement d'un modèle, la régression polynômiale n'est pas sans risque. En effet, si l'on choisit un degré de polynôme trop grand, on risque de construire une fonction qui passera par un nombre élevé de points des données d'apprentissage, mais selon une*

forme très oscillante. Une fonction oscillant trop fortement ne permettra pas de généraliser le modèle et donnera des résultats en prévision très mauvais.

2.3 Hypothèses en régression polynômiale

Comme toute autres étude et analyse lors de la proposition d'un modèle pour décrire un certain phénomène on est contraint à imposer préalablement des hypothèses sur le comportement du phénomène afin d'assurer l'adéquation du modèle proposé. Ainsi, dans le cadre de la régression polynômiale on impose ce qui suit :

- Le comportement d'une variable dépendante s'explique par une relation additive linéaire, ou curviligne, entre la variable dépendante et un ensemble de k variables indépendantes $(x_i\ , i = 1, ..., k)$.
- La relation entre la variable dépendante et toute variable indépendante est linéaire ou curviligne.
- Les variables indépendantes ne dépendent pas non plus les unes des autres.
- Les erreurs sont indépendantes, normalement distribuées avec une moyenne nulle et une variance constante.

2.4 Equation de régression polynômiale

L'équation de régression polynômiale de $n^{ème}$ degré peut s'écrire comme suit :

$$Y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \ldots + \beta_n x^n.$$

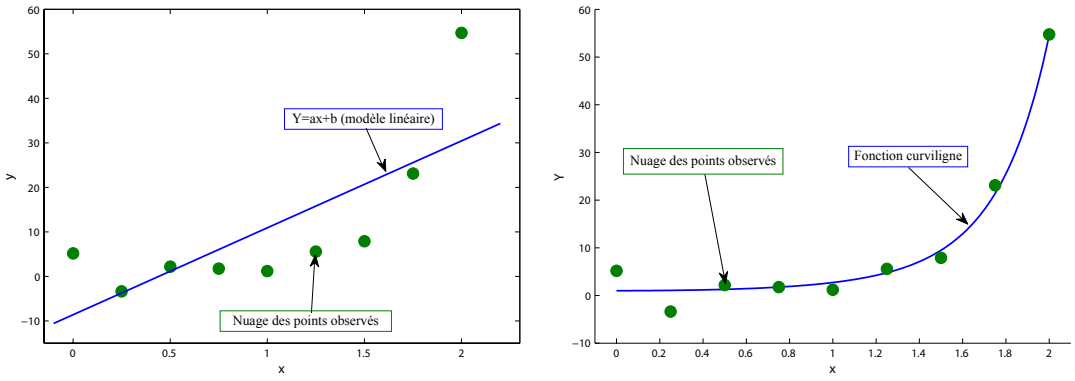
Il existe trois types de polynômes :

$$\text{Linéaire} \rightarrow \beta_0 + \beta_1 x.$$
$$\text{Quadratique} \rightarrow \beta_0 + \beta_1 x + \beta_2 x^2.$$
$$\text{Cubique} \rightarrow \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3.$$

Comme on peut le voir, le polynôme linéaire a un degré de 1, le polynôme quadratique a un degré de 2 et le polynôme cubique a un degré de 3. A mesure que le degré des équations polynômiales augmente, la courbe correspond mieux à l'ensemble de données.

Pourquoi avons-nous besoin d'une régression polynômiale ?

Le problème avec la régression linéaire était qu'elle utilise la droite du meilleur ajustement. En d'autres termes, lorsque nous avons un ensemble de données et que nous le traçons sur un graphique, il doit y avoir une droite où se trouvent les nuages de points. Mais que se passe-t-il si nous avons un ensemble de données qui ne nous donne pas de droite mais une courbe ? C'est à ce moment-là que la régression polynômiale entre en jeu.



2.5 Différence entre la régression linéaire multiple et la régression polynômiale

Voici quatre différences entre la régression linéaire multiple et la régression polynômiale :

1. **Nombre de prédicteurs :**
   - La régression linéaire multiple implique la modélisation de la relation entre une variable dépendante et deux variables indépendantes ou plus (prédicteurs).
   - La régression polynômiale, en revanche, utilise une seule variable indépendante pour prédire la variable dépendante, mais elle inclut des termes polynômiaux (par exemple, x, x^2, x^3) pour permettre des relations non linéaires.
2. **Complexité du modèle :**
   - La régression linéaire multiple suppose une relation linéaire entre la variable dépendante et chaque variable indépendante.
   - La régression polynômiale permet plus de flexibilité en ajustant une courbe aux données, ce qui peut capturer des relations non linéaires.
3. **Interprétabilité :**
   - En régression linéaire multiple, les coefficients de chaque variable indépendante représentent le changement de la variable dépendante pour un changement d'une unité dans la variable indépendante correspondante, en maintenant les autres variables constantes.
   - En régression polynômiale, l'interprétation des coefficients devient plus complexe en raison de la présence de termes polynômiaux. Chaque coefficient

correspond à l'effet du terme polynomial correspondant sur la variable dépendante.

4. **Sur-ajustement (Overfitting) :**

   - La régression polynômiale peut être sujette au sur-ajustement, surtout avec des polynômes de degré élevé, car le modèle peut capturer le bruit dans les données plutôt que la relation sous-jacente.
   - La régression linéaire multiple, lorsqu'elle est utilisée de manière appropriée, tend à être moins sujette au sur-ajustement car elle suppose une relation plus simple et linéaire entre les variables.

2.6 Estimation de β

Partant du modèle sous sa forme générale :2.1.

On souhaite construire un estimateur $\widehat{\beta} = (\widehat{\beta_0}, ..., \widehat{\beta_p})$ tel que

$$\widehat{\beta} = \arg \min_{\beta \in \mathbb{R}^{p+1}} \sum_{i=1}^{n} (Y_i - f(x_i, \beta_0, \beta_1, ..., \beta_p))^2.$$

Remarque 2.6.1 *Pour plus de détaille voir [4].*

2.7 Régression polynomial et matrices

Dans ce cas, nous voulons étudier la régression polynômiale de degré 2.

$$Y = \beta_0 + \beta_1 x + \beta_2 x^2 + \varepsilon,$$
$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \varepsilon_i, \ \{(x_i, y_i)\}\, i = 1, ..., n.$$

On a

$$MinS = \sum_{i=1}^{n} \left(\beta_0 + \beta_1 x_i + \beta_2 x_i^2 - \ y_i\right)^2 .$$

$$\frac{\partial S}{\partial \beta_0} = 0 \Rightarrow \sum_{i=1}^{n} 2\left(\beta_0 + \beta_1 x_i + \beta_2 x_i^2 - \ y_i\right) = 0 \Rightarrow \sum_{i=1}^{n} \left(\beta_0 + \beta_1 x_i + \beta_2 x_i^2\right) = \sum_{i=1}^{n} y_i \tag{2.2}$$

$$\frac{\partial S}{\partial \beta_1} = 0 \Rightarrow \sum_{i=1}^{n} 2\left(\beta_0 + \beta_1 x_i + \beta_2 x_i^2 - \ y_i\right) x_i = 0 \Rightarrow \sum_{i=1}^{n} \left(\beta_0 x_i + \beta_1 x_i^2 + \beta_2 x_i^3\right) = \sum_{i=1}^{n} y_i x_i \tag{2.3}$$

$$\frac{\partial S}{\partial \beta_2} = 0 \Rightarrow \sum_{i=1}^{n} 2\left(\beta_0 + \beta_1 x_i + \beta_2 x_i^2 - \ y_i\right) x_i^2 = 0 \Rightarrow \sum_{i=1}^{n} \left(\beta_0 x_i^2 + \beta_1 x_i^3 + \beta_2 x_i^4\right) = \sum_{i=1}^{n} y_i x_i^2 \tag{2.4}$$

1. 2.2$\Rightarrow n\beta_0 + \beta_1 \sum x_i + \beta_2 \sum x_i^2 = \sum y_i$.
2. 2.3$\Rightarrow \beta_0 \sum x_i + \beta_1 \sum x_i^2 + \beta_2 \sum x_i^3 = \sum y_i x_i$.
3. 2.4$\Rightarrow \beta_0 \sum x_i^2 + \beta_1 \sum x_i^3 + \beta_2 \sum x_i^4 = \sum y_i x_i^2$.

Alors

$$\begin{array}{ccccc} \begin{pmatrix} n & \sum x_i & \sum x_i^2 \\ \sum x_i & \sum x_i^2 & \sum x_i^3 \\ \sum x_i^2 & \sum x_i^3 & \sum x_i^4 \end{pmatrix} & \times & \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix} & = & \begin{pmatrix} \sum y_i \\ \sum y_i x_i \\ \sum y_i x_i^2 \end{pmatrix} \\ A & \times & X & = & y \end{array}$$

où $A \times X = y \Rightarrow X = yA^{-1}$.

Calculer la matrice inverse (A^{-1}) :

$$A^{-1} = \frac{1}{\det(A)} \times Com^T(A) .$$

2.8 Application numérique

Dans le présent chapitre, nous avons réalisés une étude de simulation illustrative de la mise en ouvre des modèles polynômiale pour l'approximation d'un nuage de points. Nous allons intéresser, en particulier sur l'effet de la taille de l'échantillon et de la variance des résidus ainsi que le degré du polynôme considéré dans la régression.

2.8.1 Présentation de l'application

Afin de réponde à notre objectif, cité ci-haut, nous avons conçu un programme MATLAB dont les principaux étapes sont :

Étape 1 : Générer un n-échantillon (x_i, y_i) selon un modèle cible.

Étape 2 : Estimer les paramètres du modèle polynômiale d'ordre p.

Étape 3 : Calculé l'EQM associé au régresser obtenus où

$$EQM = \frac{1}{T}\int \left(y(t) - \hat{y}(x)\right)^2 dx, \tag{2.5}$$

avec T est la plage de variation de x.

Pour les calculs numériques nous avons considéré les quatre exemples suivants :

Exemple 1 : $y = 2 + x^2$.

Exemple 2 : $y = \frac{x^3}{x+1}$.

Exemple 3 : $y = 5 * \cos(x) + x$.

Les courbes des trois modèles considérés sont présentées dans la figure 2.1.

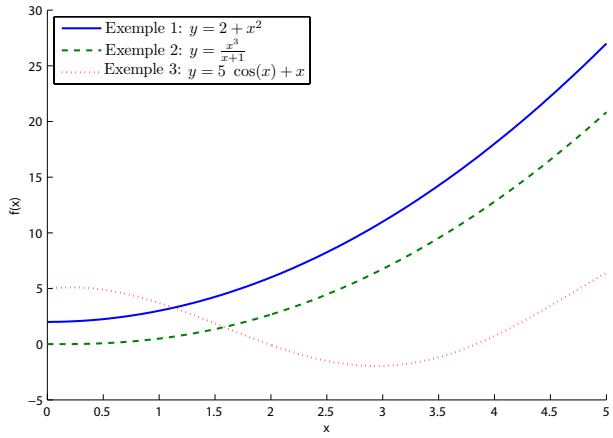


FIGURE 2.1 – Illustration graphique des Modèles considérés dans la simulation.

2.8.2 Résultats et discussion

L'exécution de notre programme MATLAB, pour une taille d'échantillon $n = 10$, $\epsilon \to N(0, 3)$ et $T = 5$ ($x \in [0, 5]$) nous a fournis les résultats numériques rangés dans le tableaux suivant et les résultats graphiques présenté dans la figure 2.2.

	$P_1(x)$	$P_2(x)$	$P_3(x)$
Exemple 1 :	9.0353	5.1628	8.5809
Exemple 2 :	8.6213	5.0593	8.5624
Exemple 3 :	10.3827	4.4550	8.6597

TABLE 2.1: EQM associés à l'approximation des données par un polynôme d'ordre 1, 2 et 3.

D'après les résultats obtenus on constate que dans les trois exemples le polynôme quadratique nous assure une meilleure approximation, au sens de l'EQM, suivis par le polynôme cubique et enfin nous trouvant la droite. Ce dernier constat coïncide parfaitement avec ce qui a été exposé dans les sections précédentes sur le modèle linéaire simple et le modèle linéaire polynômiale.

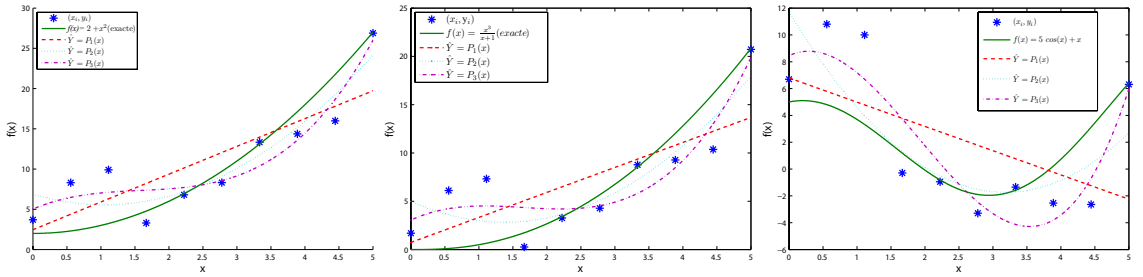


FIGURE 2.2 – Illustration graphique des régresseurs pour $p \in \{1, 2, 3\}$

L'exécution du programme, pour une autre fois pour différentes tailles d'échantillon $n \in \{10, 20, ..., 100\}$, et différentes valeurs de l'écart-type $\sigma \in \{2, 1, 0.5\}$ et $T = 5$ ($x \in [0, 5]$) lorsque nous considérons la régression par un polynôme d'ordre $p = 2$, nous a fournis les résultats présenté dans la figure 2.3 qui représente la variation de l'EQM en fonction de la taille de l'échantillon et de l'écart-type des ϵ.

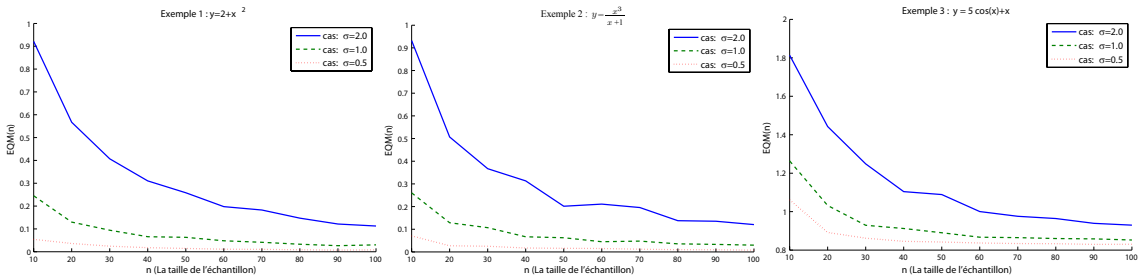


FIGURE 2.3 – Variation de l'EQM en fonction de la taille de l'échantillon et de l'écart-type des ϵ

À partir des résultats graphique présentés dans la figure 2.3 on constate que :

– En fur et à mesure que la taille de l'échantillon augmente l'erreur quadratique moyenne décroît ce qui se traduit par la pertinence du modèle proposé (à savoir le polynôme d'ordre 2) pour décrire les données issue des trois fonctions considérés.
– La qualité du modèle $\hat{y}$, au sens de l'EQM, est inversement proportionnelle à la valeur de la variance des résidus.
– La qualité du modèle $\hat{y}$, au sens de l'EQM, est proportionnelle à la taille de l'échantillon.

Conclusion

À la lumière de ce qui précède, on peut dire que la régression polynômiale représente un outil puissant et flexible pour l'analyse des données, à condition qu'elle soit utilisée avec prudence pour éviter les défis potentiels. Parmi les avantages de la régression polynômiale on a : La régression polynômiale offre une grande capacité à modéliser les relations complexes entre les variables, ce modèle est utilisé dans divers domaines tels que les sciences naturelles, ce qui en fait un outil précieux pour l'analyse des données et avec la disponibilité d'outils logiciels avancés, il est possible d'appliquer la régression polynômiale facilement et efficacement. Malgré les avantages, aucune étude n'est exempte d'inconvénients, et parmi les inconvénients de la régression polynômiale : L'augmentation du nombre de termes dans le modèle peut entraîner un ajustement excessif aux données d'entraînement, réduisant ainsi la capacité du modèle à généraliser sur de nouvelles données et l'augmentation du degré des termes peut rendre le modèle complexe et difficile à interpréter.

Bibliographie

[1] Ben ameur, S., (2022).Cours Analyse des données, Régression linéaire multiple. Université Mohamed Khider, Biskra, Algeria.

[2] Boukrif, N., (2016). Cours de Régression linéaire simple et multiple . Université Abderrahmane Mira, Bejai, Algeria.

[3] Cherfaoui, M., (2017). Cours de Biostatistiques.Statistiques Appliquées à l'expérimentation en Sciences Biologique. Université Mohamed Khider, Biskra, Algeria.

[4] Christophe, C., (2017). Modèles de régression. Cours de Master. France.
https ://cel.hal.science/cel-01248297
https ://cel.hal.science/cel-01248297v3/file/Reg-M2.pdf

[5] James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning (Vol. 112, p. 18). New York : springer.

[6] Matzner-Løber, É. (2007). Régression : Théorie et applications. Springer Science & Business Media.

[7] Introduction à la régression polynômiale : https ://delladata.fr/regression-polynômiale/.

[8] La régression linéaire simple : https ://bookdown.org/AODiakite/r4econometrics/simple-lm.html.

[9] Régression polynômiale : https ://www.studocu.com/row/u/50827160 ?sid=01716468367.

[10] Régression polynômiale : https ://www.voxco.com/fr/blog/regression-polynômiale-tout-ce-que-vous-devez-savoir/.

Annexe A : Déterminant d'une matrice

D'après les deux méthodes méthodes de régression précédentes à savoir linéaire multiple et polynômiale , on constate que le calcul des estimations des coefficients du modèle considéré se base sur le calcul matricielle en particulier le calcul de l'inverse d'une matrice. Ainsi, rappelons brièvement le calcul d'inverse d'une matrice.

Matrice carré d'ordre 2

Soit la matrice

$$M = \begin{pmatrix} a & b \\ c & d \end{pmatrix},$$

alors,

$$\det M = (a \times d) - (b \times c).$$

Matrice carré d'ordre 3

Soit la matrice

$$M = \begin{pmatrix} a^{+} & e^{-} & h^{+} \\ b & f & i \\ c & g & j \end{pmatrix}$$

alors,

$$\begin{aligned} detM &= a\begin{vmatrix} f & i \\ g & j \end{vmatrix} - e\begin{vmatrix} b & i \\ c & j \end{vmatrix} + h\begin{vmatrix} b & f \\ c & g \end{vmatrix}, \\ &= a\left|(f\times j)-(g\times i)\right| - e\left|((b\times j)-(c\times i))\right| + h\left|(b\times g)-(c\times f)\right|. \end{aligned}$$

Remarque : Une matrice M est inversible si et seulement si $\det M \neq 0$.

Calcul de la Co-matrice de M

On a

$$M = \begin{pmatrix} a & e & h \\ b & f & i \\ c & g & j \end{pmatrix}$$

et

$$com\,(M) = \begin{pmatrix} M_{11}^{+} & M_{12}^{-} & M_{13}^{+} \\ M_{21}^{-} & M_{22}^{+} & M_{23}^{-} \\ M_{31}^{+} & M_{32}^{-} & M_{33}^{+} \end{pmatrix}$$

Cofacteurs :

$$M_{11} = +\begin{vmatrix} f & i \\ g & j \end{vmatrix} = +\left|fj-ig\right|$$

$$M_{12} = -\begin{vmatrix} b & i \\ c & j \end{vmatrix} = -\left|bj-ic\right|$$

$$M_{33} = +\begin{vmatrix} a & e \\ b & f \end{vmatrix} = +\left|af-be\right|$$

Matrice$\left(M^{T}\right)$:

Soit la matrice

$$M = \begin{pmatrix} a & e & h \\ b & f & i \\ c & g & j \end{pmatrix} \Rightarrow M^T = \begin{pmatrix} a & b & c \\ e & f & g \\ h & i & j \end{pmatrix}.$$

Annexe B : Abréviations et Notations

Les différentes abréviations et notations utilisées dans ce mémoire sont expliquées ci-dessous :

y_i	:	la $i^{éme}$ observation de la variable aléatoire à expliquer$Y$.
$Cov\,(x, y)$	:	Covariance entre x et y.
ε	:	Erreur.
ddl	:	Degrés de liberté.
MCR	:	Moyenne des carrés réséduelle.
MCE	:	Moyenne des carrés expliquée.
SCT	:	La somme des carrées totale.
SCR	:	La somme des carrées résiduelle.
SCE	:	La somme des carrées expliquée.
R_q^2	:	Le coefficient de détermination du modèle réduit à $(p-q)$ variables.
$\mathcal{F}(q, n-p-1)$	:	Loi de Fisher avec $n-p-1$ ddl.
$\mathcal{T}(n-p-1)$	:	La loi de Student ave $n-p-1$ ddl.
$prév(Y)$	:	Prévision de Y.
$\frac{\partial S}{\partial \beta_0}$	:	La dérivée partielle de S par rapport à β_0.
A^{-1}	:	La matrice inverse de A.

$\det M$	:	Determinant.
$com(M)$	:	La comatrice de M.
ρ	:	Coefficient de correlation.
$Cor(x, y)$	:	Correlation entre x et y.

Résumé

La thématique abordée dans ce mémoire de Master concerne la régression polynomiale. Plus précisément, dans un premier temps, nous nous sommes intéressés à l'estimation des paramètres du polynôme en question via la méthode des moindres carrés ordinaires, où nous avons constaté que la démarche est similaire au cas de l'estimation des paramètres d'un modèle linéaire multiple. Par la suite, via une étude de simulation nous avons montré que la qualité du régresseur dépend à la fois de la taille de l'échantillon, de la variance des erreurs mais également du degré du polynôme utilisé pour modéliser les données disponibles.

Mots-clés : Régression linéaire et polynomiale, estimation paramétrique, erreurs, Méthode MCO, Simulation.

Abstract

The theme addressed in this Master's dissertation concerns polynomial regression. More precisely, initially, we were interested in estimating the parameters of the polynomial in question via the ordinary least squares method (OLS), where we noted that the approach is similar to the case of estimating the parameters of a multiple linear model. Subsequently, via a simulation studies, we showed that the quality of the regressor depends on both the size of the sample, the variance of the errors and the degree of the polynomial used to model the data.

***Key Words*:** Linear and polynomial regression, parametric estimation, errors, OLS method, simulation.

الملخص

الموضوع الذي تتناوله هذه المذكرة يتعلق بالانحدار كثير الحدود. بتعبير أدق، في البداية، كنا مهتمين بتقدير معلمات كثيرة الحدود المعنية بطريقة المربعات الصغرى العادية، حيث لاحظنا أن النهج مشابه لحالة تقدير معلمات النموذج الخطي المتعدد. بعد ذلك، من خلال دراسة محاكاة، أظهرنا أن جودة الارتداد تعتمد على حجم العينة، وتباين الأخطاء، وكذلك على درجة كثير الحدود المستخدم لنمذجة البيانات المتاحة.

الكلمات المفتاحية: الانحدار الخطي ومتعدد الحدود، التقدير، الأخطاء، طريقة أخطاء المربعات الصغرى، المحاكاة.