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 INFORMATIQUE

Polycopié du Cours

Analyse de Données

Analyse de Données en Informatique
Première Année Master (M1)

Préparé par :
Dr. AFROUN Faïrouz

Université de Biskra, 2023/2024

Table des matières

Table des figures

Chapitre 1

Régression linéaire simple et multiple

Introduction et problématique

La régression est l'une des méthodes les plus connues et les plus appliquées en statistiques pour l'analyse de données quantitatives sous forme d'un modèle. Si on s'intéresse à la relation entre deux variables, on parlera de *régression simple* en exprimant l'une des deux variables en fonction de l'autre. Tandis que, si la relation porte entre une variable et plusieurs autres variables (≥ 2), on parlera de *régression multiple*.

La mise en œuvre d'une régression impose l'existence d'une relation de cause à effet entre les variables prises en compte dans le modèle. Cette méthode peut être mise en place sur des données quantitatives observées sur n individus et présentées sous la forme :

$$y = f(x) + \epsilon, \tag{1.1}$$

où

- y est une variable quantitative prenant la valeur y_i pour l'individu i ($i = 1, ..., n$), appelée variable à expliquer ou variable réponse.
- $x_1, x_2, ..., x_p$ sont p variables quantitatives prenant respectivement les valeurs $x_{1i}, x_{2i}, ...x_{pi}$ pour le $i^{\text{ième}}$ individu, appelées variables explicatives ou prédicteurs.
- ϵ est une variable aléatoire (résidus).

Considérons un couple de variables quantitatives (X, Y). S'il existe une liaison entre ces deux variables, la connaissance de la valeur prise par X change notre incertitude concernant la réalisation de Y. Si l'on admet qu'il existe une relation de cause à effet entre X et Y, le phénomène aléatoire représenté par X peut donc servir à prédire celui représenté par Y et la liaison s'écrit sous la forme (1.1) et on dit que l'on fait de la régression de y sur x (dans le cas d'une régression multiple de y sur $x_1, x_2, ..., x_p$ la liaison peuvent être écrite sous la forme $y = f(x_1, x_2, ..., x_p)$).

Dans les cas les plus fréquents, on choisit l'ensemble des fonctions affinées du type :

Cas de régression linéaire simple :

$$f(x) = ax + b. \tag{1.2}$$

Cas de régression linéaire multiple :

$$f(x) = f(x_1, x_2, ..., x_p) = a_0 + a_1x_1 + a_2x_2 + ... + a_px_p. \tag{1.3}$$

1.1 Le modèle de régression linéaire simple

Soit un échantillon de n individus. Pour un individu i $(i = 1, ..., n)$, on a observé y_i la valeur de la réalisation de la variable quantitative Y et x_i la valeur de la variable quantitative x.

On veut étudier la relation entre ces deux variables, et en particulier, l'effet de x (variable explicative) sur y (variable réponse).

Dans un premier temps, on peut représenter graphiquement cette relation en traçant le nuage des n points de coordonnées (x_i, y_i) et on constate que la relation entre y_i et x_i s'écrit sous la forme d'un modèle de régression linéaire

La relation entre y et x est supposée n'être qu'approximative : elle est perturbée par un "terme d'erreur" additif, noté ϵ_i avec $E(\epsilon_i) = 0,\ i = \overline{1:n}$.

L'équation de la régression linéaire simple (ou le "modèle de régression") s'écrit donc de la façon suivante :

$$Y = a + bx + \epsilon, \tag{1.4}$$

$$E(Y) = a + bE(x), \tag{1.5}$$

ou encore,

$$E(Y/x) = a + bx, \tag{1.6}$$

où a et b sont les paramètres du modèle, et ϵ est le terme d'erreur qui est une variable aléatoire.

Remarque 1.1

1. *a représente le point d'intersection de la droite de régression avec l'ordonnée ("intercept", "constante").*
2. *b représente la pente de la droite de régression.*
3. *La valeur de b donne le nombre d'unités supplémentaires de Y associées à une augmentation par une unité de x.*
4. *$E(Y/x)$ est la moyenne de Y pour une valeur de x donnée.*

Exemple 1

(a) : *$Y_i = a + bx_i + cx_i^2 + \epsilon_i$, est un modèle linéaire tandis que la relation entre x et y n'est pas linéaire mais de type polynomial.*

(b) : *$Y_i = a + b\ cos(x_i) + \epsilon_i$, est un modèle linéaire.*

(c) : *$Y_i = ae^{bx_i} + \epsilon_i$, n'est pas un modèle linéaire.*

(d) : *$Y_i = ab + cx_i + \epsilon_i$, n'est pas un modèle linéaire.*

Enfin, la linéarité est relier aux paramètres du modèle et non pas aux variables explicatives.

1.2 Analyse du modèle de régression linéaire simple

Soit le couple (X, Y) de variable aléatoire où X est une variable indépendante et Y la variable dépendante. On cherche une relation du type

$$Y = a + bx + \epsilon.$$

Notons que la mise en oeuvre et l'exploitation de ce modèle nécessite une quantification préalable des paramètres inconnus a et b.

1.2.1 Estimation des paramètres du modèle

On suppose que la variable X est contrôlée par l'expérimentateur où il réalise n expériences y_1 , y_2 , $y_3, \ldots, y_n$ aux points x_1 , x_2 , x_3 , $\ldots$, x_n fixés. De plus, on suppose que les Y_i sont mutuellement indépendants.

Le modèle s'écrit

$$y_i = a + bx_i$$

pour $i = \overline{1:n}$, tel que :

- $E(\epsilon_i) = 0$
- $Cov(\epsilon_i, \epsilon_j) = 0 \quad \forall i \neq j$
- $Var(\epsilon_i) = \sigma^2 \quad \forall i = \overline{1:n}$

Supposons qu'on opte pour la méthode des moindre carrées pour quantifier a et b, alors les estimateurs des paramètres a et b sont $\hat{a}$ et $\hat{b}$ qui minimise la fonction $Q(a, b)$, définie par :

$$Q(a,b) = \sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} (Y_i - a - bx)^2. \tag{1.7}$$

Cela revient à la détermination d'un optimum minimal de la fonction des erreurs quadratique $Q(a, b)$, qui consiste à résoudre le système des équations suivant :

$$\begin{cases} \frac{\partial Q(a,b)}{\partial a} &= 0, \\ \frac{\partial Q(a,b)}{\partial b} &= 0, \end{cases} \tag{1.8}$$

c'est-à-dire,

$$\begin{cases} -2\sum_{i=1}^{n}(Y_i - a - bx) &= 0 \\ -2\sum_{i=1}^{n} x_i(Y_i - a - bx) &= 0 \end{cases} \Rightarrow \begin{cases} \sum_{i=1}^{n} Y_i - \sum_{i=1}^{n} a - b\sum_{i=1}^{n} x_i &= 0 \\ \sum_{i=1}^{n} Y_i x_i - a\sum_{i=1}^{n} x_i - b\sum_{i=1}^{n} x_i^2 &= 0 \end{cases}. \tag{1.9}$$

Finalement, le système à résoudre, pour estimer les coefficients de régression a et b, ni rien d'autre qu'un système linéaire à deux équations et à deux inconnus, qui est donné par :

$$\begin{cases} a\left(\sum_{i=1}^{n} 1\right) + b\left(\sum_{i=1}^{n} x_i\right) &= \sum_{i=1}^{n} Y_i \\ a\left(\sum_{i=1}^{n} x_i\right) + b\left(\sum_{i=1}^{n} x_i^2\right) &= \sum_{i=1}^{n} Y_i x_i \end{cases} \tag{1.10}$$

La résolution du système (1.10), nous fournis la solution suivante :

$$\boxed{\hat{b} = \frac{\frac{1}{n}\sum_{i=n}^{n} x_i y_i - \overline{X}\ \ \overline{Y}}{\frac{1}{n}\sum_{i=1}^{n} x_i^2 - \overline{X}^2}} \text{ et } \boxed{\hat{a} = \frac{1}{n}\sum_{i=1}^{n} Y_i - \hat{b}\frac{1}{n}\sum_{i=1}^{n} x_i}, \tag{1.11}$$

ou encore :

$$\boxed{\hat{b} = \frac{Cov(x,y)}{Var(x)}} \text{ et } \boxed{\hat{a} = \overline{Y} - \hat{b}\overline{X},} \tag{1.12}$$

où :

$$Cov(x,y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{X})(y_i - \overline{Y}) = \left[\frac{1}{n}\sum_{i=1}^{n} x_i y_i\right] - \overline{X}\ \overline{Y}. \tag{1.13}$$

1.2.2 Estimation de σ^2

En plus de l'estimation des paramètres du modèle (a et b), l'une des caractéristique statistique importante liée au modèle est bien que la variance inconnue σ^2. Pour cela, nous allons estimer σ^2, où nous proposons d'utiliser la méthode MLE (voir chapitre 2). À cet effet, on suppose que $\epsilon_i \rightsquigarrow \mathcal{N}(0, \sigma^2)$, alors

$$Y_i \rightsquigarrow \mathcal{N}(a + bx_i, \sigma^2).$$

Dans ce cas, la fonction de vraisemblance correspondante au modèle est donnée par :

$$\mathcal{L}(Y_1, Y_2, ..., Y_n, \sigma^2) = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \exp\left[-\frac{1}{2\sigma}\sum_{i=1}^{n}(y_i - \hat{a} - \hat{b}x_i)^2\right],$$

L'expression de la variance qui maximise cette fonction est donnée par :

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{a} - \hat{b}x_i)^2. \tag{1.14}$$

Mais, cet estimateur est un estimateur avec Biais, alors il doit être corrigé. Ainsi, après sa correction on aura l'estimateur sans Biais de σ^2 suivant :

$$\hat{\sigma}_c^2 = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{a} - \hat{b}x_i)^2 \rightsquigarrow \chi^2_{n-2}. \tag{1.15}$$

1.2.3 Qualité et validation du modèle :

Dans cette section, nous allons présenter deux manière du jugé la qualité et l'adéquation du modèle linéaire :

$$Y_i = a + bx_i + \epsilon_i \quad , i = 1, ..., n.$$

pour l'explication de la variable Y à l'aide de la variable x.

1.2.3.1 Coefficients de corrélation et 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 variables.

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 :

$$\rho = Cor(x, y) = \frac{Cov(x, y)}{\sigma_x \sigma_y} = r(x, y). \tag{1.16}$$

avec,

$$\sigma_x^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{X})^2, \tag{1.17}$$

$$\sigma_y^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \overline{Y})^2, \tag{1.18}$$

et $Cov(x, y)$ est donnée dans (1.13).

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.

Le coefficient de corrélation nous donne des informations sur l'existence d'une relation linéaire entre les deux variables considérées. À cet effet, un coefficient de corrélation nul ne signifie pas l'absence de toute relation entre les deux variables mais seulement l'absence d'une relation linéaire. Pour cela, il ne faut pas confondre la corrélation et la relation causale. Une bonne corrélation entre deux variables peut révéler une relation de cause à effet entre elle, mais pas nécessairement.

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 :

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

Ce nombre mesure l'adéquation entre le modèle et les données observées où plus, R^2 est près de 1, plus le modèle est adéquat et le contraire est vrai.

1.2.3.2 Le test de Fisher

Une autre technique, plus puissante que le calcule de coefficient de corrélation, pour mesuré la pertinence et l'adéquation d'un modèle est l'utilisation du test de Ficher qui se base sur l'analyse de la variance.

On peut démontrer que la variation totale de Y se décompose comme suit :

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

où

SCT : Variation de Y ou variation totale.

SCR : Variation des résidus.

SCE : Variation de la régression ou variation expliqué par la régression.

Dans la logique des choses, le modèle est validé, si la variation totale du modèle n'est engendré que par la variation des résidus et non pas par la variation de la régression, autrement dit la variation moyenne des résidus doit être supérieur à la variation moyenne de la régression pour valider le modèle,

$$\frac{\sum_{i=1}^{n}(\hat{y}_i - \overline{Y})^2/1}{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2/(n-2)} > k > 1,$$

donc, il nous reste à savoir comment déterminer la valeur critique k.

Sachant que :

$$\frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \rightsquigarrow \chi^2_{n-2}$$

et

$$\frac{1}{\sigma^2}\sum_{i=1}^{n}(\hat{y}_i - \overline{Y})^2 \rightsquigarrow \chi^2_1,$$

alors,

$$F = \frac{\sum_{i=1}^{n}(\hat{y}_i - \overline{Y})^2/1}{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2/(n-2)} \rightsquigarrow f_{(1,n-2)}$$

où la notation $f_{(1,n-2)}$ désigne est une loi de Fisher de degrés de liberté $n_1 = 1$ et $n_2 = n-2$, cela signifie que, pour un risque α, la valeur critique k n'est rien d'autre que le fractale d'ordre $1-\alpha$ d'une loi de Fisher de degrés de liberté 1 et $n-1$ ($k = f_{(1,n-2,1-\alpha)}$) ainsi on décide que :

- Si $f_c > f_{(1,n-2,1-\alpha)}$ alors le modèle est valide.
- Si $f_c \leq f_{(1,n-2,1-\alpha)}$ le modèle n'est pas valide.

où f_c est la réalisation de la statistique F.

1.3 Régression linéaire multiple

Dans la pratique les principales étapes d'une analyse de régression Multiple sont :

1. Définir la variable dépendante et les variables explicatives.
2. Spécifier la nature de la relation entre la variable dépendante et les variables explicatives.
3. Estimer les paramètres du modèle, en suite, quantifier sa qualité et vérifier sa validité.
4. Dans le cas où le modèle est retenu, interpréter sa signification par rapport au problème posé.

Dans cette section, nous nous intéresserons à la régression multiple dans le cadre du modèle linéaire. La régression linéaire multiple est la généralisation de la régression linéaire simple qui ne considère qu'une seule variable explicative. Considérons le modèle linéaire multiple dont la forme est la suivante :

$$Y = b_1 x_1 + b_2 x_2 + ... + b_k x_k + \epsilon,$$

pour la $i^{\text{ième}}$ observation le modèle peut être représenté de la manière suivante :

$$Y_i = b_1 x_{1i} + b_2 x_{2i} + ... + b_k x_{ki} + \epsilon_i \ , \qquad i = 1, ..., n,$$

ou encore, sous sa forme Matricielle :

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} x_{11} & x_{21} & ... & x_{k1} \\ x_{12} & x_{22} & ... & x_{k2} \\ \vdots & & & \vdots \\ x_{1n} & x_{2n} & ... & x_{kn} \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_3 \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}.$$

$$Y \quad = \quad X \quad b \quad + \quad \epsilon$$

À partir des étapes de l'analyse de régression multiple, cité précédemment, on est au niveau de l'étape (3), c'est-à-dire on doit estimer les paramètres (coefficients) du modèle.

1.3.1 Estimation des paramètres du modèle

Supposons qu'on a :

$$Y = Xb + \epsilon,$$

avec,

$$X = \begin{bmatrix} x_{11} & x_{21} & ... & x_{k1} \\ x_{12} & x_{22} & ... & x_{k2} \\ \vdots & & & \vdots \\ x_{1n} & x_{2n} & ... & x_{kn} \end{bmatrix}, \quad b = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{bmatrix}, \quad Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \quad et \quad \epsilon = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

de plus,

- $E(\epsilon) = 0$,
- $Var(\epsilon) = \sigma^2 I_n$, où I_n est une matrice d'identités d'ordre n.

On utilisant la méthode des moindres carrés, pour estimer les coefficients du modèle, on aura un système linéaire à k équations et k variables. Ce système s'écrit sous sa forme matricielle comme suit :

$$\underbrace{\begin{bmatrix} \sum_{i=1}^{n} x_{1i}^2 & \sum_{i=1}^{n} x_{1i}x_{2i} & ... & \sum_{i=1}^{n} x_{1i}x_{ki} \\ \sum_{i=1}^{n} x_{1i}x_{2i} & \sum_{i=1}^{n} x_{1i}^2 & ... & \sum_{i=1}^{n} x_{2i}x_{ki} \\ \vdots & & & \vdots \\ \sum_{i=1}^{n} x_{1i}x_{ki} & ... & ... & \sum_{i=1}^{n} x_{ki}^2 \end{bmatrix}}_{M} \times \underbrace{\begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_k \end{bmatrix}}_{b} = \underbrace{\begin{bmatrix} \sum_{i=1}^{n} y_i x_{1i} \\ \sum_{i=1}^{n} y_i x_{2i} \\ \vdots \\ \sum_{i=1}^{n} y_i x_{ki} \end{bmatrix}}_{m}$$

où $M = X^t\ X$ et $m = X^t\ Y$.

Finalement, l'estimation des coefficients du modèle sont données par le calcul matriciel suivant :

$$\hat{b} = (X^t.X)^{-1}.X^t.Y.$$

1.3.2 Test sur la validité du modèle

Avec le même raisonnement abordé dans le cas de la régression linéaire simple on peut construire le test de validation du modèle. En effet, la variation totale de Y se décompose comme suit :

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

où

SCT : Variation de Y ou variation totale.

SCR : Variation des résidus.

SCE : Variation de la régression ou variation expliqué par la régression.

Pour valider le modèle, on test

$$H_0 \text{ "}b_1 = b_2 = ... = b_k = 0\text{" contre } H_1 \text{ "}\exists j \in \{1,\ 2, ...,\ k\}/\ b_j \neq 0\text{"},$$

avec le même raisonnement que dans le cas de régression linéaire on obtient la statistique, du test, suivante :

$$F = \frac{\sum_{i=1}^{n}(\hat{y}_i - \overline{Y})^2/(k-1)}{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2/(n-k)} \rightsquigarrow f_{(k-1,n-k)},$$

où la notation $f_{(k-1,n-k)}$ désigne la loi de Fisher à $k-1$ et $n-k$ degrés de liberté.

Ainsi, on décide :

- Si $f_c > f_{(k-1,n-k,1-\alpha)} \Rightarrow$ de valider le modèle.
- Si $f_c \leq f_{(k-1,n-k,1-\alpha)} \Rightarrow$ de ne pas valider le modèle.

avec f_c est la réalisation de la statistique F.