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 Biologie

Module Biostatistique Master 1
Chapitre 03
Regréssion Linéaire Simple :

Par
Prof : CHALA ADEL

2021-2022

Je dédie ce travail............
A mes parents ils m'ont tous,
avec leurs moyens, soutenu et donné
la force d'aller toujours
plus loin.
A ma chère femme Houda.
A l'esprit du professeur Bahlali Seid

Table des matières

Introduction

En Master 1, les étudiants de Sciences de la Nature et de la Vie se voient proposer des formations spécialisées nécessitant des connaissances et un savoir-faire statistiques qui ne peuvent être acquis en tronc commun. D'autre part, certains étudiants, sans envisager a priori une spécialisation statistique, peuvent désirer acquérir une formation approfondie en méthodes statistiques.

Cette formation est particulièrement appréciée pour un débouché professionnel dans les domaines de l'expérimentation, et préparation de Master et post Doctoral.

Chaque méthode statistique est motivée par une présentation de problèmes concrets, par des utilisateurs dans différents domaines : agronomie, écologie, génétique, médecine, ...

Les connaissances acquises concernent l'estimation des paramètres, les tests statistiques (validité du modèle, effet des variables explicatives), la prévision et la sélection de variables dans le cadre du modèle linéaire (régression simple, analyse de la variance à plusieurs facteurs, analyse de la covariance).

Choix du modèle en fonction du type de données, structuration des données, traitement statistique et informatique (logiciel SPSS) pour les modèles de régression multiple et d'analyse de la variance ou de la covariance, ainsi que pour des extensions de ce modèle (ACP, AFC).

Ce module apporte une formation solide en statistique inférentielle directement exploitable dans de nombreux Masters Végétale ou Master Biologie Moléculaires, ou options de Poste Doctoral. Il donne des compétences indispensables pour la collecte et le traitement de données expérimentales. En cela, il constitue un pré requis important pour des formations en génétique, écologie et en sciences de l'environnement (Master Biologie)

Mots clés : Statistique inférentielle, modèle linéaire (régression, analyse de la variance et de la covariance), modèles mixtes (effet aléatoire), sélection de variables

Chapitre 1

Régression linéaire simple

Soit une distribution à deux variables quantitatives. La régression linéaire simple permet de chercher l'éventuelle relation fonctionnelle linéaire qui existerait entre une valeur EXPLICATIVE (ou indépendante) x et une variable aléatoire À EXPLIQUER (ou dépendante) y, x est remplacé par t s'il s'agit d'une mesure du temps.

Graphiquement, on représente cette éventuelle relation dans un plan muni d'un repère orthogonal. L'axe des abscisses indique la variable qui explique et l'axe des ordonnées celle que l'on cherche à expliquer. L'ensemble des données figure sous forme de nuage de points (autant de points que d'observations différentes). Si les données sont disponibles en fourchettes de valeurs, on remplace ces dernières par les valeurs centrales des classes.

Une relation linéaire déterministe entre les deux variables se traduit par des points parfaitement alignés. En mathématiques, on dit que la droite qui les relie représente une fonction affine (en statistiques, on emploie un peu abusivement le terme LINÉAIRE plutôt qu'AFFINE). Toutefois, on ne trouve jamais de relation parfaite en utilisant des données brutes, sauf à vérifier une définition (auquel cas on parle de modèle déterministe). La relation est donc stochastique, c'est-à-dire qu'elle comporte une part d'aléas.

La régression linéaire simple cherche à modéliser cette relation par une équation et l'analyse de corrélation vise à en évaluer la qualité.

Ce type d'analyse peut d'ailleurs être utilisé pour des relations non linéaires mais transformables en fonctions affines à condition d'utiliser des variables auxiliaires (voir régression simple sur tendance exponentielle). En pratique, il est toutefois rare de passer par là puisque n'importe quel logiciel réalise des régressions non linéaires.

On cherche à approximer une liaison inconnue $Y = f(X)$ par une relation linéaire suivante

$$Y_i = \alpha X_i + \beta + \xi_i, \qquad \text{où } i = \overline{1, n},$$

où ξ_i étant une va riable aléatoire normale, la variable Y est la variable aléatoire expliquer, de plus $\mathrm{E}(\xi_i) = 0$, et $Var(\xi_i) = \sigma^2$.

1.1 Estimation des paramètres α et β :

Le modèle d'estimation est celui des moindres carrés ordinaires (MCO) et consiste à trouver les estimateurs $\widehat{\alpha}$ et $\widehat{\beta}$ de α et β (resp) qui minimisent la somme des carrés des erreurs suivante :

$$Q = \sum_1^n \xi_i^2 = \sum_1^n (y_i - \alpha x_i - \beta)^2 .$$

La résolution de $\min Q$ conduit à donner les deux estimateurs pour α et β qui sont :

$$\begin{aligned} \widehat{\alpha} &= \frac{\sum_1^n (x_i - \overline{x})(y_i - \overline{y})}{\sum_1^n (x_i - \overline{x})^2}, \\ \widehat{\beta} &= \overline{y}_n - \alpha \overline{x}_n. \end{aligned}$$

Remarque : En désignant par $S_X^2 = \frac{1}{n}\sum_1^n (x_i - \overline{x})^2$ et $S_Y^2 = \frac{1}{n}\sum_1^n (y_i - \overline{y})^2$ les variances empiriques de $(x_1,\ x_2,\ x_3, ...,\ x_n)$ et $(y_1, y_2, y_3, ..., y_n)$ et par ρ le coefficient de corrélation linéaire empirique entre X et Y donné par :

$$\begin{aligned} \rho &= \frac{\sum_1^n (x_i - \overline{x})(y_i - \overline{y})}{n S_X^2 S_Y^2} = \frac{S_{XY}}{S_X^2 S_Y^2}, \\ \text{on a } \widehat{\alpha} &= \rho \frac{S_Y}{S_X}. \end{aligned}$$

1.1.1 La droite de régression

Nous observons un nuage de forme plus ou moins rectiligne. Comment trouver l'équation de la droite qui le résume au mieux ? En minimisant les distances qui la séparent des points. Quelles distances ? Généralement les carrés des distances euclidiennes parce que l'utilisation des valeurs absolues nous bloquerait dans une impasse mathématique un peu longue à expliquer (mais certains logiciels permettent de réaliser ce type de régression). D'où l'expression droite des moindres carrés. Graphiquement, il s'agit des distances VERTICALES, parallèles à l'axe y, la distance entre le modèle théorique et la réalité.

1.2 La qualité de l'ajustement :

On appelle :

$$\widehat{y}_i = \widehat{\alpha} x_i + \widehat{\beta} : \text{ La valeur ajustée de } y_i.$$

$$y = \widehat{\alpha} x + \widehat{\beta} : \text{ La droite de régression linéaire de } y \text{ ajustée en } x.$$

$$\widehat{\xi}_i = y_i - \widehat{y}_i : \text{Le résidu en } i.$$

$$V_E = \frac{1}{n} \sum_1^n \left(\widehat{y}_i - \overline{y}_n\right)^2 : \text{ La variance expliquée par le modèle.}$$

$$V_R = \frac{1}{n} \sum_1^n \widehat{\xi}_i^2 : \text{ La variance résiduelle.}$$

On montre que :

$$V_T = \frac{1}{n} \sum_1^n \left(y_i - \overline{y}\right)^2 = V_E + V_R : \text{ L'équation d'analyse de la variance.}$$

La qualité du modèle est jugée par le coefficient de détermination de la régression R^2 :

$$\begin{aligned} R^2 &= \frac{V_E}{V_T} = \frac{\left(\widehat{\alpha}\right)^2 \sum_1^n \left(x_i - \overline{x}\right)^2}{\sum_1^n \left(y_i - \overline{y}\right)^2} \\ &= 1 - \frac{\sum_1^n \widehat{\xi}_i^2}{\sum_1^n \left(y_i - \overline{y}\right)^2} = \rho^2. \end{aligned}$$

1.2.1 Interprétation

Si le coefficient de corrélation est suffisamment élevé, le modèle peut-être utilisé pour des applications prédictives ou prévisionnelles. On remplace alors l'inconnue x dans l'équation de la droite et l'on obtient une estimation de l'ordonnée qui lui correspond. En général, on procède à une extrapolation : graphiquement, on prolonge la droite.

Toutefois, le modèle peut être meilleur si l'équation d'une courbe remplace celle de la droite (lorsque le nuage de points présente une forme de banane), ou en ajoutant une deuxième variable explicative. Outre la connaissance « métier » du sujet, c'est l'observation des résidus qui doit mettre la puce à l'oreille (voir hypothèses de validité de la régression linéaire).

1.3 Propriétés Statistiques de $\widehat{\alpha}$ et $\widehat{\beta}$:

Les estimateurs $\widehat{\alpha}$ et $\widehat{\beta}$ sont sans biais, alors on peut calculer leurs espérance et la variance comme suit :

$$E\left(\widehat{\alpha}\right) = \alpha.$$

$$E\left(\widehat{\beta}\right) = \beta.$$

$$Var\left(\widehat{\beta}\right) = \sigma^2 \left(\frac{1}{n} + \frac{\left(\overline{x}\right)^2}{\sum_1^n \left(x_i - \overline{x}\right)^2}\right).$$

$$Var\left(\widehat{\alpha}\right) = \frac{\sigma^2}{\sum_1^n \left(x_i - \overline{x}\right)^2}.$$

Un estimateur sans biais de σ^2 c'est :

$$S^2 = \frac{\sum_1^n \widehat{\xi}_i^2}{n-2}.$$

1.4 Intervalle de confiance pour α :

En utilisant les techniques des statistiques, on peut établir l'intervalle de confiance suivant :

$$P\left(\widehat{\alpha} - t_a^* \frac{S}{\sqrt{\sum_1^n \left(x_i - \overline{x}\right)^2}} < \alpha < \widehat{\alpha} + t_a^* \frac{S}{\sqrt{\sum_1^n \left(x_i - \overline{x}\right)^2}}\right) = 1 - a,$$

où t_a^* est la fractile d'ordre $\left(1 - \frac{a}{2}\right)$ pour la loi de Student a $(n-2)$ DDL.

1.5 Test de Student

On veut tester l'hypothèse nulle suivante $H_0 : \left(\widehat{\alpha} = \alpha_0\right)$ contre l'hypothèse alternative suivante $H_1 : \left(\widehat{\alpha} \neq \alpha_0\right)$, pour cela on doit utiliser la loi de Student pour aboutir la formule d'estimation qui reste vraie et acceptable pour tout valeur t_a est la fractile d'ordre $\left(1-\frac{a}{2}\right)$ pour la loi de Student à $(n-2)$ DDL :

$$\mathcal{T}=\frac{|\widehat{\alpha}-\alpha_0|}{\frac{S}{\sqrt{\sum_1^n (x_i-\overline{x})^2}}}<t_a$$

Exemple N°: 01

On a étudié les longueurs respectives des 2 (deux) paires d'ailes d'une espèce de guêpe (Vespa sp) sur un échantillon de 11 individus. Soit X la longueur d'une aile de la première paire et Y celle de l'aile de la deuxième paire mesurée sur le même individu. On a obtenu les résultats suivants :

L'individu	La longueur d'une aile de la première paire	La longueur d'une aile de la deuxièm
I	294	624
II	271	661
III	314	728
IV	356	782
V	383	819
V I	369	869
V II	402	938
V III	422	1023
IX	475	1136
X	475	1227
XI	486	1317

On veut tester la réalité d'une relation linéaire entre Y et X, soit :

$$Y=\beta+\alpha X+\xi.$$

Les hypothèses classiques de modèle linéaire simple sont supposées réalisées, c'est-à-dire que ξ est une variable aléatoire suivant la loi normale centrée.

1/ Donner les estimations $\widehat{\alpha}$ et $\widehat{\beta}$ de α et β obtenues par la méthode des moindres carrés ordinaires MCO.

2/ Calculer l'estimateur de la variance, l'estimateur de la variance de l'estimateur $\widehat{\beta}$, et le coefficient de détermination de la régression R^2.

3/ On se pose le problème de test :

$$\begin{cases} H_0 : \widehat{\alpha}=0, \\ H_1 : \widehat{\alpha}\neq 0. \end{cases}$$

A quelle question ce test répond t-il ? Peut-on dire que α est significativement différent de zéro, au risque $0,05$?

La réponse :

$\sum_1^{11} x_i = 10124$	$\sum_1^{11} y_i^2 = 9850614,000$	$\sum_1^{11} x_i y_i = 4075178$
$\sum_1^{11} y_i = 4247$	$\overline{x} = 386,091$	$\sum_1^{n} (x_i - \overline{x})^2 = 55464,137$
$\sum_1^{11} x_i^2 = 1695193,000$	$\overline{y} = 920,364$	$\sum_1^{n} (y_i - \overline{y})^2 = 532845,183$

1/ **Les estimateurs $\widehat{\alpha}$ et $\widehat{\beta}$ de α et β :**

$$\widehat{\alpha} = \frac{\sum x_i y_i - n(\overline{x})(\overline{y})}{\sum x_i^2 - n(\overline{x})^2} = 3,000.$$

$$\widehat{\beta} = \overline{y}_n - \alpha \overline{x}_n = -237,909$$

L'estimateur $\mathbf{S}^2$ de σ^2 :

$$S^2 = \widehat{\sigma}^2 = \frac{\sum_1^n \widehat{\xi}_i^2}{n-2} = \frac{\sum_1^n (y_i - \overline{y})^2 - (\widehat{\alpha})^2 \sum_1^n (x_i - \overline{x})^2}{n-2} = 3740,883.$$

Alors $S = 19,98$. Donc :

$$Var(\widehat{\alpha}) = \frac{S^2}{\sum_1^n (x_i - \overline{x})^2} = 0,067.$$

Coefficient de détermination de la régression $\mathbf{R}^2$:

$$\begin{aligned} R^2 &= \rho^2 = 1 - \frac{\sum_1^n \widehat{\xi}_i^2}{\sum_1^n (y_i - \overline{y})^2} \\ &= 1 - \quad = 0,93. \end{aligned}$$

Test de Student :

On pose le problème de test suivant

$$\begin{cases} H_0 : \widehat{\alpha} = 0, \\ H_1 : \widehat{\alpha} \neq 0. \end{cases}$$

pour cela on doit utiliser la loi de Student pour aboutir la formule d'estimation qui reste vraie et acceptable pour tout valeur t_a est la fractile d'ordre $\left(1 - \frac{a}{2}\right)$ pour la loi de Student a $(n-2)$ DDL :

$$\mathcal{T} = \frac{|\widehat{\alpha} - \alpha_0|}{\frac{S}{\sqrt{\sum_1^n (x_i - \overline{x})^2}}} = \frac{|0,312|}{\frac{19,98}{\sqrt{\sum_1^n (x_i - \overline{x})^2}}} = 11,39 > t_a = t^{n-2}_{\left(1-\frac{\alpha}{2}\right)} = 2,262.$$

Le coefficient $\widehat{\alpha}$ est très significativement différent de 0.

Exemple N°: 02

Pour vérifier les relations d'halométrie entre insectes, on a retenu les deux relations x La longueur de l'élytre y La largeur de la tête. Les mesures sur 50 insectes, notées $(x_i; y_i)$ ont fourni les résultats suivants :

$$\begin{aligned}\sum_1^{50} x_i &= 155, \quad \sum_1^{50} y_i = 125, \quad \sum_1^{50} x_i y_i = 391,1.\\ \sum_1^{50} x_i^2 &= 482,5, \quad \sum_1^{50} y_i^2 = 320,5, \quad \sum_1^{50} x_i^2 y_i^2 = 3468,7.\end{aligned}$$

1/ Calculer :

a/ La moyenne et l'écart-type du caractère x sur l'échantillon observée.

b/ La moyenne et l'écart-type du caractère y sur l'échantillon observée.

c/ La covariance empirique et la corrélation empirique des variables x et y.

d/ L'équation de la droite de régression linéaire de y sur x obtenue par estimation sur ces données.

2/ Donner un estimateur de la variance pour l'estimateur du coefficient.

3/ Déterminer l'intervalle de confiance pour le coefficient $\widehat{\alpha}$ au taux de confiance 95%.

La réponse :

Pour $n = 50$.

1/ a) et b) Calcul des moyennes :

$$\overline{x} = \frac{1}{n}\sum_1^{50} x_i = \frac{1}{50}.155 = 3,1.$$

$$\overline{y} = \frac{1}{n}\sum_1^{50} y_i = \frac{1}{50}.125 = 2,5.$$

Calcul des variances :

$$\sigma_X = \sqrt{VarX} = \sqrt{\frac{1}{n}\sum_1^{n} (x_i - \overline{x})^2} = \sqrt{\frac{1}{n}\sum_1^{50} x_i^2 - (\overline{x})^2} = 0,2.$$

$$\sigma_Y = \sqrt{VarY} = \sqrt{\frac{1}{n}\sum_{1}^{n}(y_i - \overline{y})^2} = \sqrt{\frac{1}{n}\sum_{1}^{50} y_i^2 - (\overline{y})^2} = 0,4.$$

c) La covariance :

$$\begin{aligned} S_{XY} &= Cov(X,Y) = \frac{1}{n}\sum_{1}^{50} x_i y_i - (\overline{x})(\overline{y}) \\ &= \frac{1}{50}391,1 - (3,1)(2,5) = 0,072. \end{aligned}$$

et le coefficient de corrélation ρ :

$$\rho = \frac{S_{XY}}{S_X S_Y} = \frac{0,072}{0,2 \times 0,4} = 0,9.$$

d) Les estimateurs des coefficients :

$$\widehat{\alpha} = \rho \frac{S_Y}{S_X} = 0,9\frac{0,4}{0,2} = 1,8.$$

$$\widehat{\beta} = \overline{y}_n - \alpha \overline{x}_n = -3,081.$$

Alors la droite de la régression linéaire obtenue par la méthode des moindre carré ordinaire est la suivante

$$y = 1,8x - 3,081$$

2/ L'estimateur de la variance pour l'estimateur du coefficient $\widehat{\alpha}$: on a la variance de l'estimateur $\widehat{\alpha}$ est donnée par

$$Var(\widehat{\alpha}) = \frac{\sigma^2}{\sum_1^n (x_i - \overline{x})^2}.$$

Donc :

$$\widehat{Var(\widehat{\alpha})} = \frac{S^2}{\sum_1^n (x_i - \overline{x})^2} = \frac{\frac{\sum_1^n \widehat{\xi}_i^2}{n-2}}{\sum_1^n (x_i - \overline{x})^2} = \frac{\frac{\sum_1^n (y_i - \overline{y})^2 - (\widehat{\alpha})^2 \sum_1^n (x_i - \overline{x})^2}{n-2}}{\sum_1^n (x_i - \overline{x})^2} = 0,015.$$

Chapitre 2

Exercice sur La régression Linéaire simple

Exercice N°: 01

On considère le modèle linéaire :

$$y_t = \alpha x_t + \beta + \varepsilon_t$$

Où xi et yi représentent respectivement le nombre d'heures d'ensoleillement et le montant des ventes de crèmes glacées pour une région donnée et durant le mois de juin de l'année t. On suppose que les hypothèses assurant l'interprétation probabiliste de la méthode des MCO :

T	1	2	3	4	5
X	330	309	279	312	318
Y	32,4	32,1	29,7	31,2	31,1

1/ préciser le modèle.

2/ Donner une estimateur de $V(\alpha)$.

3/ Tester si l'ensoleillement liée significativement les ventes.

Exercice N°: 02

Les œufs d'une espèce particulière d'insecte nécessitants avant leur développement un séjour à basse température(Hivernation). On étudie la corrélation éventuelle qui liée la durée d'hivernation x et le temps de développement de ces œufs(x et y en jours).

1/ Calculer les estimation pour α et β à partir des observations suivantes :

$\overline{x} = 1,8$, $\overline{y} = 1,7$, $s_X^2 = \frac{1}{10}\sum_1^{10}(x_i - \overline{x})^2 = 1,9$, $s_Y^2 = \frac{1}{10}\sum_1^{10}(y_i - \overline{y})^2 = 26,7$, $s_{XY} = \frac{1}{10}\sum_1^{10}(x_i - \overline{x})(y_i - \overline{y}) = -6,2.$

2/ La variable X a-t-elle une influence significative sur Y au niveau $0,01$si $\widehat{\sigma(\widehat{a})} = 0,446$?

Exercice N°: 03

On désire avoir s'il existe une relation entre le poids de naissance d'un enfant et âge de sa mère à l'accouchement. Dans ce but, on prélève 100 dossiers médicaux dans le fichier des naissances d'une maternité. On calcule les quantités suivantes :

	Poids de naissance de l'enfant	L'age de la mère à l'accouchement
Moyenne observée	3100	25
Variance observée S^2	10 000	25

La covariance observée entre le poids de naissance de l'enfant et l'âge de la mère à l'accouchement :

$$s_{XY} = \frac{1}{n}\sum_1^n (x_i - \overline{x})(y_i - \overline{y}) = 450 \text{ gr/ans}$$

1/ Donner la valeur numérique du coefficient de corrélation entre le poids de naissance de l'enfant et l'âge de sa mère à l'accouchement.

2/ Calculer la droite de régression linéaire entre le poids de naissance de l'enfant et l'âge de sa mère à l'accouchement.

3/ La qualité du modèle est jugée par le coefficient de détermination de la régression, exprimer leur valeur.

4/ Au risque de 5% le poids de naissance de l'enfant et l'âge de sa mère à l'accouchement sont-ils significativement liés ?

Exercice N :04

D'avril 2008 à juillet 2010 on a relevé les données relatives aux variables suivantes pour le mois t :

y_t : Ecart entre les versements et les retraits dans les Caisses CASNOS.

x_{1t} : Accroissement relatif du revenu des ménages.

x_{2t} : Accroissement relatif de l'indice des prix de détail.

On a obtenu :

$$\begin{aligned}\sum y_t &= 39,82; \qquad \sum x_{1t} = 25,14; \qquad \sum y_t^2 = 63,21;\\ \sum_t^2 x_{1t}^2 &= 24,36; \qquad \sum x_{1t}y_t = 38,38.\end{aligned}$$

1/ Ecrire l'équation de la droite de régression de y_t en x_{1t} et donner un indicateur de la qualité de cette régression.

2/ On effectue maintenant la régression y_t en x_{2t}, le calcul des résidus d'ajustement donne $\sum \varepsilon_i^2 = 1,96$. Comparer avec le modèle précédent.

Exercice N :05

Le revenue R_t et l'épargne nette E_t ont été mesurés par trimestre pendant 3 ans pour une catégorie socioprofessionnelle bien déterminée ; Après correction des variations saisonnières, exprimées en million de Dinars, les indicateurs suivants sont disponibles :

$$\begin{aligned}\overline{R}_t &= \frac{1}{12}\sum_1^{12} R_i = 19,7, \qquad \sum_1^{12} R_i^2 = 4827,\\ \overline{E}_t &= \frac{1}{12}\sum_1^{12} E_i = 6,1, \qquad \sum_1^{12} E_i^2 = 456.\end{aligned}$$

$$\sum_1^{12} R_i E_i = 1480$$

On suppose que les variables E_t et R_t sont liée par le modèle :

$$E_t = \beta + \alpha R_t + u_t$$

Les v,a étant de la loi $\mathcal{N}(0;2)$ pour tout t et indépendantes.

1/ Calculer $\widehat{\alpha}$ et $\widehat{\beta}$, estimateurs des MCO de α et β.

2/ Etudier la validité du modèle.

3/ On désire tester l'hypothèse qu'une augmentation absolue de 1% du revenu implique une augmentation absolue de 1% de l'épargne. Ecrire cette hypothèse en fonction des coefficients de la régression et résoudre le problème.

.

Bibliographie

[1] G. Saporta ; Probabilités, analyse des données et statistique. Technip, Paris, 1990.

[2] B. Escofier, J. Pagés ; Analyse factorielles simple et multiple. Dunod, Paris, 1998.

[3] Avner Bar-Hen : Cours de DEUG Probabilités et Statistiques, Université Aix-Marseille III, 2002–2003.

[4] Admane, O., Hoang Ky., Ouakli N : Statistique (cours et exercices) Pour les étudiants du tranc commun bio-médical., OPU 1998.

[5] M Vilain : Méthodes expérimentales en agronomie- Pratique et analyse. Edition Tec et Doc. 1999.