الجمهوريـة الجزائـريـة الديمقراطيـة الشعبيـة
République Algérienne Démocratique et Populaire
وزارة التعليـم العالـي والبحـث العلمـي
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

N° Réf :……

Centre Universitaire

Abd elhafid Boussouf Mila

Institut des sciences et de la technologie **Département de Mathématiques et Informatiques**

Mémoire préparé En vue de l'obtention du diplôme de Master

En: Mathématiques

Spécialité : Mathématiques fondamentales et appliquées

La régression Linéaire Dans Les Essais Cliniques

Préparé par :

- **Bahloul Ramla**
- **Kebbout Khaoula**

Soutenue devant le jury

-Encadré par : A.Zerari ……. M.A.B

-Président : W.Fadal …… M.A.B

-Examinateur : C.Sekhane ……. M.A.B

Année universitaire : 2015/2016

Table des matières

Remerciements

Ce mémoire n'aurait pu se réaliser sans des nombreux soutiens qui m'ont été apportés au cours de sa concrétisation.

Dieu soit loué, c'est à grâce à son appui que je suis arrivée à ce stade.

Je remercie et de tout profond du cœur madame Zerari Amel, maitre assistant classe B, Centre Universitaire Abdelhafid Boussouf Mila, qui mérite mes insères reconnaissances d'avoir accepté de m'encadrer.

Je remercie ma sœur LABDAOUI Ahlem pour votre aide, immolassions, et orientement.

Mille merci nos trop chères familles, que dieu vous garde pour nous.

Nous tenons à exprimer, nos sincères remerciements à tout les enseignants qui nous ont formé durant les années d'étude, et tous ceux qui nous ont apporté une aide au pour la réalisation de ce projet.

Sans oublier bien-sûre tous les amis et collègues d'études pour leur enjouement et soutient moral.

Dédicace

Je dédie ce mémoire à ma mère "Zahiya", qui m'a encouragé à aller de l'avant et qui m'a donné tout son amour pour reprendre mes études.
A mon père "Zitouni " pour le soutien qu'il m'a toujours donne.
A mes frères :"Kayes", "Bilal", "Salim" et ses enfants :"Meryem", "Adem", "Fouzi" et "Konouz".
A ma soeur "Soulaf" et ses enfants :"Aya", "Bessmala", "Ishak", et "Abdelali".
A mon oncle "Hani", et ses enfants :"Chouki", et "Abdelwahab".
Je tiens à d´dédier ce travail à mes chères amis:"Khawla", "Bouthaina", "Dounia", et spécialement mon binôme et mon sœur "Khaoula".

RAMLA

Dédicace

A mes parents humble témoignage de ma profonde affection et de ma reconnaissance pour tous les sacrifices qu'ils ont consentis

A mes sœurs

A ma famille

A tous les amis et mon binôme

Khaoula

Introduction Générale

Le progrès de la thérapeutique a été de tout temps un des soucis principaux du médecin, mais les moyens employés sont longtemps restés dans le domaine de l'empirisme, des impressions, de l'à peu près. Bien entendu, la thérapeutique a quand même progressé.
Les progrès en thérapeutique sont de plus en plus souvent constitués par une succession de gains modestes, parfois même on recherche des traitements non pas plus efficaces, mais mieux tolérés. Dans ce domaine des différences fines, seul peut apporter une réponse un essai rigoureusement mené.
C'est depuis une époque récente, pratiquement la fin de la dernière guerre mondiale, qu'a été élaborée une méthodologie véritablement scientifique, conférant aux essais la rigueur réservée jusqu'alors aux expériences de laboratoire, et ceci essentiellement par l'intervention de la méthode statistique.
Ce mémoire a été motivé par une application des mathématiques et la statistique en particulier dans les essais cliniques.
A ce propos, nous avons considéré trois chapitres :
Dans le premier chapitre, nous avons définis les essais cliniques, leurs déférentes phases, leurs objectifs et la conduite pratique qui contient la mise en œuvre d'un protocole et le questionnaire.
Dans le deuxième chapitre, nous avons définis théoriquement la régression linéaire simple et multiple avec interprétations.
Enfin, dans le troisième chapitre, nous avons faire un application de la régession lineaire sur les essias clinique avec déffirentes méthodes.

Chapitre 1

Les essais cliniques

Introduction

La recherche clinique est une activité médicale visant à améliorer la connaissance soit d'une maladie, soit d'une thérapeutique. La recherche clinique concerne l'être humain.
En pharmacologie,la recherche clinique est dominée par les études du médicament administré à l'homme dans le cadre des essais cliniques(essais cliniques ou évaluations cliniques).
Les essais cliniques se déroulent après les études dites de pharmacologie expérimentale,qui se déroulent en laboratoire(stade préclinique). Les études chez l'homme obéissent à une technique(méthodologie des essais cliniques), à une législation et à une éthique.Ces études peuvent se dérouler soit en médecine de ville, soit dans les centres hospitaliers,soit dans des structures de recherche agrées publiques ou privées.

1.1 Objectifs

1.1.1 Pourquoi les essais cliniques sont-ils nécessaires ?

Avant de proposer de nouveaux traitements aux patients concernés, il faut s'assurer qu'ils sont efficaces et bien tolérés. Les essais cliniques permettent non seulement de valider de nouveaux traitements mais aussi de définir pour quelles populations de patients ils sont les plus efficaces.
Enfin, ils permettent de mieux comprendre les caractéristiques d'une maladie. Les essais cliniques sont obligatoires dans la procédure permettant la mise sur le marché d'un médicament.

1.1.2 Sur quels critères un essai clinique est-il décidé ?

Si on reconsidère par exemple l'étude de l'identification d'une molécule en recherche, celle-ci est évaluée en différentes étapes au laboratoire puis sur l'animal. Ces étapes permettent d'avoir une première évaluation de sa tolérance et de son intérêt thérapeutique. Si ces données sont satisfaisantes, les tests sur l'homme sont envisagés. Le laboratoire pharmaceutique dépose alors une demande auprès d'un Comité d'Ethique Indépendant, chargé de revoir l'ensemble du protocole et du déroulement de l'essai. Le comité donne un avis en particulier sur la pertinence du projet et la protection des personnes qui vont y participer.
L'essai clinique ne débutera qu'après avis favorable du comité, qui suivra régulièrement l'avancée de l'essai.

1.1.3 Quelle est l'importance des essais cliniques ?

Les essais cliniques font la démonstration de ce qui fonctionne ou non en médecine.ils apportent des réponses à d'importantes questions que les scientifiques se posent et font progresser la recherche.
Pour le passé, c'est grâce à des essais cliniques que les médecins ont pu mettre au point de nouvelles techniques chirurgicales mieux tolérées par les patients, élaborer de nouveaux médicaments qui traitent plus efficacement certains types de maladie et trouver des traitements qui entraînent moins d'effets secondaires.

1.1.4 Quels sont les inconvénients de la participation à un essai clinique ?

Les traitements administrés aux sujets qui participent à un essai clinique étant de nature expérimentale, ils peuvent ne pas être efficaces ou causer des effets indésirables désagréables, voire dangereux. Les sujets qui prennent part à des essais cliniques doivent suivre très attentivement les instructions des médecins.
La participation à un essai clinique peut aussi exiger du temps. Le patient qui participe à un essai passe plus de temps en examen et en évaluation qu'un patient ordinaire.

1.2 Définitions

1.2.1 Qu'un ce qu'un essai clinique ?

Un essai clinique est une recherche biomédicale organisée et pratiquée sur l'homme en vue du développement des connaissances biologiques ou médicales.
L'essai peut se faire sur un volontaire malade ou un volontaire sain.
Pour débuter l'essai doit avoir obtenu un avis favorable du Comité de protection des personnes (CPP) et une autorisation de l'Agence nationale de sécurité du médicament et des produits de santé (ANSM).
L'essai clinique cherche essentiellement à répondre à deux questions :

Comment est-il tolèré ?	Le médicament est-il éfficace ?
Si un médicament est actif, il est par définition potentiellement dangereux et sa mauvaise utilisation peut, selon les cas : - soit provoquer des troubles relativement mineurs (vertiges, nausées, sécheresse de la bouche...) - soit provoquer des accidents éventuellement graves (syncope, troubles du rythme cardiaque, anémie, hémorragie ...) Certes, un nouveau médicament n'est utilisé chez l'homme qu'après que des études de toxicologie et de pharmacologie menées sur des animaux de laboratoire aient permis de préciser la dose utile, le meilleur mode d'administration et les principales précautions à prendre. Il n'en demeure pas moins que l'Homme n'est ni un rat ni un lapin et que tous les êtres humains ne réagissent pas de la même façon. Il est donc indispensable de confirmer la bonne tolérance et la sécurité du médicament dans ses conditions normales d'utilisation, en particulier quand il est associé à d'autres traitements (risque d'interactions médicamenteuses).	Il ne suffit pas de constater la disparition d'un symptôme pour affirmer que le médicament est efficace. Cette amélioration ou guérison peut en effet être dûe à d'autres causes, par exemple : - l'évolution naturelle d'une maladie dont l'organisme est capable de se guérir tout seul. exemples : Un rhume, une entorse bénigne... - à l'influence d'un autre facteur thérapeutique, qui n'est pas obligatoirement un médicament exemple : vous vous couchez en ayant mal à la tête et le lendemain, au réveil, ce mal de tête a disparu, probablement sous l'effet du repos. - à un effet placebo, c'est à dire un effet psychologique favorable qui peut faire disparaitre certains symptômes ou au moins les faire passer au second plan. exemple : Encore votre mal de tête .... que vous avez presque oublié pendant toute la journée à cause de votre travail mais qui devient lancinant le soir, au moment où vous vous couchez. Le travail et vos soucis ont eu un effet placebo. Ils n'ont pas soigné votre mal de tête (ils l'ont même peut être aggravé) mais grâce à eux, vous n'y pensiez plus et donc le ressentiez moins.

1.2.2 Les différentes phases d'un essai clinique

Phase I :
Lors de la phase I, les essais sont, généralement, réalisés chez le volontaire sain (c'est-à-dire non malade).
Ces essais ont lieu dans des centres spécialisés qui ont reçu un agrément de la part des autorités de santé.
Ces études ont deux objectifs majeurs :
Premièrement, il s'agit de s'assurer que les résultats concernant la toxicité obtenus lors du développement préclinique, sont comparables à ceux obtenus chez l'homme. Cela permet de déterminer quelle est la dose maximale du médicament en développement tolérée chez l'homme.
Deuxièmement, il s'agit de mesurer, via des études de pharmacocinétique, le devenir du médicament au sein de l'organisme en fonction de son mode d'administration (absorption, diffusion, métabolisme et excrétion).

Phase II :
Les essais sont réalisés sur des patients. Leur objectif est de tester l'efficacité du produit et de déterminer la dose optimale (posologie).
Ces études sont le plus souvent comparatives : l'un des 2 groupes de patients reçoit la molécule tandis que l'autre reçoit un placebo.

Phase III :
Ces essais, de plus grande envergure, sont conduits sur plusieurs milliers de patients représentatifs de la population de malades à laquelle le traitement est destiné.
Il s'agit d'essais comparatifs au cours desquels le médicament en développement est comparé à un traitement efficace déjà commercialisé ou, dans certains cas, à un placebo, c'est-à-dire un traitement sans activité pharmacologique.
Cette comparaison se fait, le plus souvent, en double insu et avec tirage au sort, c'est-à-dire que les traitements sont attribués de manière aléatoire sans que le patient et le médecin chargé du suivi soient informés de quelle attribution ils ont fait l'objet.
Ces essais visent à démontrer l'intérêt thérapeutique du médicament et à en évaluer son rapport bénéfice/risque.
C'est à l'issue de la phase III que les résultats peuvent être soumis aux Autorités Européennes de Santé pour l'obtention de l'autorisation de commercialisation appelée AMM (Autorisation de Mise sur le Marché).

Phase IV :
Les essais de phase IV sont réalisés une fois le médicament commercialisé, sur un nombre de patients souvent très important (jusqu'à plusieurs dizaines de milliers de personnes).
Ils permettent d'approfondir la connaissance du médicament dans les conditions réelles d'utilisation et d'évaluer à grande échelle sa tolérance.
La pharmacovigilance permet ainsi de détecter des effets indésirables très rares qui n'ont pu être mis en évidence lors des autres phases d'essai.

1.2.3 Effet placebo

L'effet d'un traitement peut être indépendant de son activité chimique ou physique. Il peut aussi être du à la pensée qu'a le patient de son traitement (relation malade-traitement) et à la confiance qu'il porte à son médecin (relation médecin-malade). La guérison (ou pas) d'un patient peut donc être influencée par des conditions psychologiques particulières (penser qu'on va guérir aide à la guérison, le contraire est aussi vrai).
Un médicament placebo est un comprimé sans aucune activité chimique ou physique.
Il est facile de faire un médicament placebo. C'est un peu plus compliqué pour d'autres études, comme la chirurgie ou des interventions cliniques.

1.2.4 Conduite d'un essai clinique

Pour déterminer la valeur relative du nouveau traitement par rapport à un traitement standard (ou un placebo).
Un échantillon de patients ayant la maladie est constitué,Cet échantillon est séparé en deux groupes de patients (théoriquement, ces deux groupes sont en tous points semblables).
L'un des groupes est soumis au nouveau traitement, l'autre au traitement standard ou au placebo.
Les groupes sont comparés quant à l'effet désiré.

1.2.5 Quels sont les principaux acteurs dans un essai clinique ?

Le promoteur :
Il s'agit de la personne ou l'institution qui a l'initiative d'un essai.
Il s'agit dans la plupart des cas d'une entreprise du médicament, mais il peut s'agir d'hôpitaux ou de centres de recherche (INSERM, c'est-à-dire l'Institut National de la Santé Et de la Recherche Médicale).

Les investigateurs :
Un investigateur est un médecin et aussi un spécialiste expérimenté de recherche clinique qui prépare un protocole ou un plan de traitement dans le cadre d'un essai clinique et qui le réalise chez des malades.

Les patients :
Pour participer à un essai, les patients doivent répondre à des critères très précis.
Ceci permet de sélectionner les sujets pour lesquels le traitement étudié sera le mieux adapté, et d'exclure ceux pour lesquels le traitement est contre-indiqué.

1.3 Critère spécifiques des essais cliniques

1.3.1 Comment les patients sont-ils informés ?

Une recherche ne peut être menée sans information de la personne sur laquelle est mené l'essai et sans qu'elle ait donné son consentement libre et éclairé. Avant d'accepter ou de ne pas accepter de participer à un essai clinique, la personne est informée par le médecin qui dirige l'essai, le médecin investigateur ou un médecin le représentant. L'information doit être objective, loyale et compréhensible par le sujet. Toutes ces données sont résumées dans un document d'information écrit remis à la personne dont le consentement est sollicité. Le CPP (Comité de Protection des Personnes) donne son avis sur ces documents.

1.3.2 Comment les patients sont-ils protégés ?

Dans le cadre des essais cliniques, l'information des sujets ou des patients que l'on se propose d'inclure dans les essais est un élément capital de la protection de ces personnes, validé par un CPP (Comité de Protection des Personnes). Tout patient susceptible de s'engager dans un protocole d'essai clinique doit signer un document dit de « consentement éclairé » qui garantit qu'il a reçu de la part du médecin investigateur (ou d'un médecin qui le remplace) toutes les informations concernant :
- Les objectifs, la méthodologie et la durée de la recherche.
- Les bénéfices attendus de la recherche.
- Les contraintes et les risques prévisibles, y compris en cas d'arrêt de la recherche avant son terme.
- Des éventuelles alternatives médicales.
- La prise en charge médicale en fin de recherche si nécessaire, ou en cas d'arrêt prématuré ou d'exclusion de la recherche.
- L'avis du CPP (Comité de Protection des Personnes) et l'autorisation de l'autorité compétente.
- Si besoin, l'interdiction de participer simultanément à une autre recherche et/ou la période d'exclusion qui suit la recherche ainsi que l'inscription du participant dans le fichier national.
- Le droit au refus de participer.
- La possibilité de retrait du consentement à tout moment sans encourir aucune responsabilité, ni aucun préjudice.
- La communication au participant des informations concernant sa santé au cours ou à l'issue de la recherche.
- L'information sur les résultats globaux de la recherche à la fin de l'essai selon des modalités qui sont précisées dans ce document.

1.3.3 Avantages et risques des essais cliniques pour le patient

Il n'est pas nécessairement facile de décider de participer ou non à un essai clinique .Vous devez toutefois garder à l'esprit que ces études sont conçues dans le but d'obtenir les meilleurs résultats possible tout en limitant les risques pour les participants.Il est important de discuter avec votre médecin des avantages et risques spécifiques de tout essai clinique .

Parmi les avantages possibles :
- Vous pourriez profiter d'un traitement qui n'est pas offert autrement et qui pourrait s'avérer plus sécuritaire ou efficace que les options actuelles.
- Même si vous ne faites pas partie du groupe qui reçoit le nouveau traitement standard disponible.
- En vous renseignant sur les essais cliniques et les options thérapeutiques, vous participez activement à une décision qui a un effet direct sur votre vie.
- Vous avez l'occasion de venir en aide à d'autres personnes et d'enrichir les connaissances sur la maladie.

Parmi les risques possibles :
- Les nouveaux traitements se sont pas toujours meilleurs ni aussi efficaces que les traitements standards .
- Des effets secondaires imprévus pourraient se manifester.
- Le nouveau traitement, même s'il est valable, pourrait ne pas fonctionner pour vous.
- Si vous faites partie du groupe qui reçoit le traitement standard, il se peut que vous n'obteniez pas d'aussi bons résultats que les participants qui recevront le nouveau traitement.
- La participation à un essai clinique peut imposer des contraintes ou demander plus de temps.
- Vous devrez peut-être subir des examens supplémentaires ou prendre plus de médicaments.

1.3.4 Qui peut participer à un essai clinique ?

Les essais cliniques comportent des ensembles de règles bien établies (appelés protocoles) définissant les personnes aptes à y participer. Ces protocoles exposent généralement l'état de santé exact attendu des participants et peuvent préciser d'autres conditions, notamment en matière d'âge, de sexe, etc. Par exemple, l'essai clinique d'un nouveau traitement de l'hypertension artérielle peut recruter des sujets de plus de 50 ans souffrant d'hypertension modérée. Si vous remplissez les conditions exposées dans les protocoles, vous seriez admissible à participer à l'essai. Les règles de participation à un essai clinique sont très rigoureuses, pour plusieurs raisons : par souci de sécurité, les chercheurs doivent recueillir des renseignements très spécifiques sur les sujets

qui présentent un certain type de pathologie et d'antécédents médicaux, ils souhaitent aussi recruter des personnes qui seront vraisemblablement aidées par le nouveau traitement et qui présentent le risque le plus faible possible d'effets indésirables.
Si vous souhaitez participer à un essai clinique, demandez d'abord conseil à votre médecin de famille. Il pourra vous orienter vers un essai clinique en cours dans un hôpital ou un centre de recherche de votre région.

1.3.5 Où se déroulent les essais cliniques ?

Les essais cliniques se déroulent dans des centres d'investigation clinique implantés au sein d'établissements hospitaliers, ou dans des centres privés d'investigation. Les lieux ou sont réalisées les études préliminaires doivent être autorisés et sont régulièrement inspectés.

1.4 La démarche expérimentale

1.4.1 Essai à l'aveugle

Qu'il s'agisse de l'égalité dans l'appréciation ou de l'égalité dans l'évolution de la maladie, il apparait que les essais « simplement à l'aveugle » et mieux encore « doublement à l'aveugle » apportent les garanties maximum de rigueur. Mais il va de soi qu'ils se heurtent en contrepartie à des difficultés d'ordre éthique ou matériel, voire à des impossibilités. Par ailleurs, on a bien l'idée que la conduite « à l'aveugle », indispensable dans l'étude de soporifiques ou d'antalgiques, pourrait être évitée dans certains essais, sur le cancer par exemple. Mais alors cette lacune devrait être présente à l'esprit au moment de l'interprétation critique des résultats. Il s'agit finalement dans chaque cas de peser le pour et le contre.

Simple aveugle :
Dans ce cas votre médecin connait le traitement qui vous est donné mais pas vous. Vous prenez un médicament sans savoir si c'est A, B ou un placebo.
Pourquoi ? Parce que le fait de savoir ce que vous prenez peut influencer de façon importante votre jugement et donc vos réponses aux questions posées par le médecin qui vous suit dans cet essai.

Double aveugle :
Ni vous ni votre médecin ne savent si vous prenez le médicament A et B ou un placebo.
Pourquoi ? Parce que comme vous, votre médecin peut avoir un jugement faussé, ou au moins un préjugé s'il doit analyser les effets du traitement en sachant qu'il s'agit d'un médicament actif ou d'un placebo. Pour la qualité d'une étude, il est préférable que le médecin soit aussi « aveugle » que vous pour ne pas introduire de subjectivité dans sa façon de conduire un examen, de vous interroger et de consigner les résultats.
Mais c'est fou !et s'il m'arrive quelque chose, qui saura quoi faire ?
Rassurez vous, tout le monde n'est pas aveugle dans ce type d'essai. Un centre vigilance, accessible $24h/24$ et $7j/7$, sait exactement qui prend quoi. En cas de besoin, votre médecin peut interroger ce centre pour « lever l'aveugle » ou « lever l'anonymat » et prendre alors toutes les mesures adaptées.
Il faut néanmoins savoir que le fait de « lever l'aveugle » pour un patient conduit le plus souvent à interrompre l'essai chez celui-ci (mais pas obligatoirement les consultations de surveillance éventuellement prévues). Votre médecin ne peut donc « lever l'aveugle » par simple curiosité (la votre ou la sienne).
A l'inverse, à la fin d'une étude, il est habituel de lever l'aveugle et vous pourrez demander à votre médecin ce que vous preniez. Vous aurez peut être une surprise ! Certains patients, et leurs médecins, sont ravis de l'efficacité du placebo. A l'inverse, dans prés de la moitié des levées d'aveugle faites en cours d'essais en raison d'effets secondaires considérés comme suffisamment préoccupants pour faire interrompre l'étude, on découvre que le patient était en fait sous placebo ! Eh, oui, l'homme n'est pas une machine....et cela explique que les études en phase III voire IV soient très fréquemment réalisées en double aveugle.

1.5 Le protocole

Un protocole est un plan d'étude spécifique à chaque essai clinique. Ce plan est soigneusement élaboré, aussi bien pour garantir la santé des participant, que pour apporter des réponses aux questions identifiées au début de l'essai.
Un protocole décrit le type d'individus, les procédures, les critères d'évaluation, les médicaments et leur posologie, ainsi que la durée de l'étude. Pendant qu'ils

participent à un essai clinique, les participent sont régulièrement examinés par le personnel chargé de la recherche, qui surveille leur état de santé et détermine l'efficacité de leur traitement.
Les codes d'éthique et juridiques qui gouvernent la pratique médicale s'appliquent plus spécifiquement aux essais cliniques.
Les essais suivent un protocole soigneusement contrôlé, un plan qui détaille ce que les chercheurs feront au cours de l'étude. Au fur et à mesure qu'un essai clinique progresse, les chercheurs rapportent les résultats de l'essai aux différentes agences gouvernementales. Les noms des individus participant resteront secrets et ne seront pas mentionnés dans ces rapports. L'orsqu' ils sont terminés, leur résultat est mise en ligne par les entreprises du médicament, qu'ils soient positifs ou négatifs.
Tout essai clinique doit être approuvé au préalable et contrôlé par un Comité d'Ethique (appelé Comité Institutionnel de Contrôle , ou par des Comités d'Ethique Indépendants (CEI) dans l'Union Européenne), afin de garantir que les risques seront aussi faibles que possible par rapport aux bénéfices attenus et vérifiés.
Un Comité d'Ethique est un Comité indépendant constitué de médecins, de statisticiens, de juristes indépendant, de représentants de la société civile (profanes en la matière) et d'autres experts dument qualifiés qui garantissent qu'un essai clinique est éthique et que les droits des participant à l'étude sont protégés.

1.5.1 La mise en œuvre d'un protocole

1.5.1.1 Sujets bons pour l'essai

A. La maladie :
- Définition de la maladie .
- Formes (cliniques, histologiques...).
B. Les malades :
- Clauses générales (sexe, âge...).
- Lieu de recrutement.
- Traitements antérieurs.
- Contre-indication, clause d'ambivalence.
- Elimination pour abandon probable du traitement.
- Elimination pour surveillance difficile.
- Clauses particulières.

L'ensemble A et B doit être réalisé par deux séries de clauses :
- D'abord des critères d'inclusion dans l'essai, définissant une catégorie assez large de malades.
- Ensuite les clauses d'exclusion, éliminant des sujets de la catégorie ci-dessus.

1.5.1.2 Les traitements

Chacun des traitements est défini dans ses grandes lignes. On précise ensuite dans le plus grand détail :
- les conditions d'administration.
- le cotexte (autres traitements principaux, traitements accessoires, régime. . .).
- point important : l'essai est –il « simplement à l'aveugle » ou « doublement à l'aveugle » ?.

1.5.1.3 Le tirage au sort Préciser

- le moment.
- le procédé adopté.

1.5.1.4 Critères de jugement

Leur détermination dans les grandes lignes (nature explicative ou pragmatique, un ou plusieurs critères).
Critères retenus, leur mode de meure.
Moment de la mesure. Préciser très exactement le protocole de surveillance.

1.5.1.5 Y a-t-il un plan expérimental ?

Si oui, lequel ?

1.5.1.6 Analyse classique ou progressive ?

1.5.1.7 Nombre de sujets nécessaire :

Expliquer comment il a été déterminé. Eventualité de l'essai inter-centres.

1.5.1.8 Organisation générale de l'essai :

Désignation du coordinateur et des responsables dans chaque service.
Modalités de liaison entre les divers participants.
Evaluation de la durée probable de l'essai.

Chapitre 2

La régression linéaire

Introduction

L'origine du mot régression vient de Sir Francis Galton. En 1885, travaillant sur l'hérédité, il chercha à expliquer la taille des fils en fonction de celle des pères. Il constata que lorsque le père était plus grand que la moyenne, « taller Than mediocrity », son fils avait tendance à être plus petit que lui ET, a contrario, que lorsque le père était plus petit que la moyenne, « shorter than mediocrity », son fils avait tendance à être plus grand que lui. Ces résultats l'ont conduit à considérer sa théorie de « regression toward mediocrity ». Cependant l'analyse de causalité entre plusieurs variables est plus ancienne et remonte au milieu du xviii siècle.
En 1757, R. Boscovich, né à Ragussa, l'actuelle Dubrovnik, proposa une méthode minimisant la somme des valeurs absolues entre un modèle de causalité et les observations. Ensuite Legendre dans son célèbre article de 1805, « Nouvelles méthodes pour la détermination des orbites des comètes », introduit la méthode d'estimation par moindres carrés des coefficients d'un modèle de causalité et donna le nom à la méthode. Parallèlement, Gauss publia en 1809 un travail sur le mouvement des corps célestes qui contenait un développement de la méthode des moindres carrés, qu'il affirmait utiliser depuis 1795 (Birkes et Dodge, 1993).
Dans ce chapitre, nous allons analyser la régression linéaire simple : nous pouvons la voir comme une technique statistique permettant de modéliser la relation linéaire entre une variable explicative (notée X) et une variable à expliquer (notée Y). Cette présentation va nous permettre d'exposer la régression linéaire dans un cas simple afin de bien comprendre les enjeux de cette méthode, les problèmes posés et les réponses apportées.

2.1 Où se place la régression linéaire ?

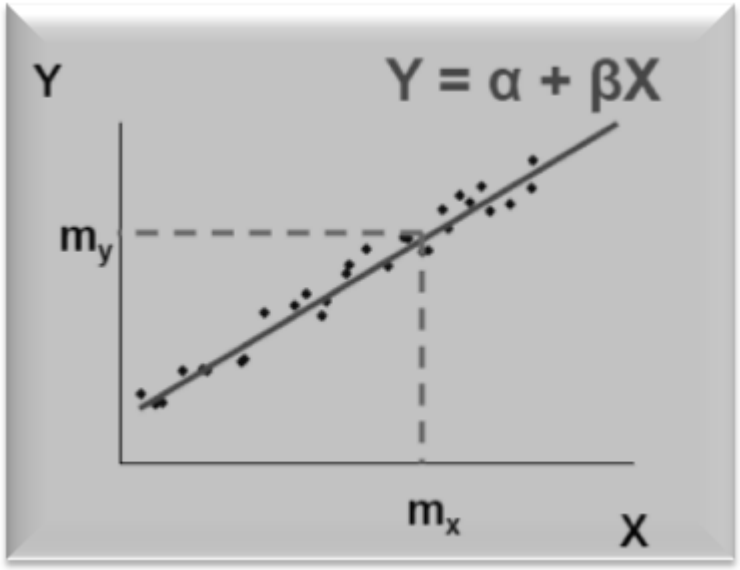
La régression linéaire se classe parmi les méthodes d'analyses multivariées qui traitent des données quantitatives.
C'est une méthode d'investigation sur données d'observations, ou d'expérimentations, où l'objectif principal est de rechercher une liaison linéaire entre une variable Y quantitative et une ou plusieurs variables X également quantitatives.
C'est la méthode la plus utilisée pour deux raisons majeures :

1. c'est une méthode ancienne,
2. c'est l'outil de base de la plupart des modélisations plus sophistiquées comme la régression logistique, le modèle linéaire généralisé, les méthodes de traitement des séries temporelles, et surtout des modèles économétriques, etc.

L'analyse de régression se décompose en plusieurs étapes : énonce du problème, sélection des variables potentiellement pertinentes, collecte des données, spécification du modèle, choix de la méthode d'ajustement, ajustement du modèle, validation et utilisation du modèle choisi pour la solution du problème pose.

2.2 La régression linéaire simple

La régression s'adresse à un type de problème où les 2 variables quantitatives continues X et Y ont un rôle asymétrique : la variable Y dépend de la variable X.
La liaison entre la variable Y dépendante et la variable X indépendante peut être modélisée par une fonction de type $Y = \alpha + \beta X$, représentée graphiquement par une droite :



$Y = \alpha + \beta X$
Y : variable dépendante (expliquée)
X : variable indépendante (explicative)
α : ordonnée à l'origine (valeur de Y pour $x = 0$)
β : pente (variation moyenne de la valeur de Y pour une augmentation d'une unité de X)

2.2.1 Modélisation statistique

Lorsque nous ajustons par une droite les données, nous supposons implicitement qu'elles étaient de la forme :

$$Y = \beta_1 + \beta_2 X$$

Mais puisque cette liaison est perturbée par un « bruit ». Nous supposons en fait que les données suivent le modèle suivant :

$$Y = \beta_1 + \beta_2 X + \varepsilon \tag{2.1}$$

L'équation (2.1) est appelée modèle de régression linéaire et dans ce cas précis modèle de régression linéaire simple. Les β_j, appelés les paramètres du modèle (constante de régression et coefficient de régression), sont fixes mais inconnus, et nous voulons les estimer. La quantité notée ε est appelée bruit, ou erreur, et est aléatoire et inconnue.

2.2.2 Moindres Carrés

Les points (x_i, y_i) étant donnés, le but est maintenant de trouver une fonction affine f telle que la quantité $\sum_{i=1}^{n} L(y_i - f(x_i))$ soit minimale. Pour pouvoir déterminer f, encore faut-il préciser la fonction de coût L. Deux fonctions sont classiquement utilisées :

- coût absolu $L(u) = |u|$;
- le coût quadratique $L(u) = u^2$.

Les deux ont leurs vertus, mais on privilégiera dans la suite la fonction de coût quadratique. On parle alors de méthode d'estimation par moindres carrés (terminologie due à Legendre dans un article de 1805 sur la détermination des orbites des comètes).

2.2.2.1 Estimateurs des moindres carrés

Définition 1 *On appelle estimateurs des moindres carrés (en abrégé MC) $\hat{\beta}_1$ et $\hat{\beta}_2$ les valeurs minimisant la quantité :*

$$S(\beta_1, \beta_2) = \sum_{i=1}^{n} (y_i - \beta_1 - \beta_2 x_i)^2.$$

Autrement dit, la droite des moindres carrés minimise la somme des carrés des distances verticales des points (x_i, y_i) du nuage à la droite ajustée :

$$Y = \hat{\beta}_1 + \hat{\beta}_2 X$$

2.2.2.2 Calcul des estimateurs de β_1 et β_2

La fonction de deux variables S est une fonction quadratique et sa minimisation ne pose aucun problème, comme nous allons le voir maintenant.

Proposition 1 *(Estimateurs $\hat{\beta}_1$ et $\hat{\beta}_2$)*
Les estimateurs des MC ont pour expressions :

$$\hat{\beta}_1 = \bar{Y} - \hat{\beta}_2 \bar{X}$$

avec :

$$\hat{\beta}_2 = \frac{\sum_{i=1}^{n} (x_i - \bar{X})(y_i - \bar{Y})}{\sum_{i=1}^{n} (x_i - \bar{X})(x_i - \bar{X})} = \frac{\sum_{i=1}^{n} (x_i - \bar{X}) y_i}{\sum_{i=1}^{n} (x_i - \bar{X})^2}$$

Preuve 1 *La première méthode consiste à remarquer que la fonction $S(\beta_1, \beta_2)$ est strictement convexe, donc qu'elle admet un minimum en un unique point $(\hat{\beta}_1, \hat{\beta}_2)$, lequel est déterminé en annulant les dérivées partielles de S. On obtient les "équations normales" :*

$$\begin{cases} \dfrac{\partial S}{\partial \beta_1} &= -2\sum_{i=1}^{n}(y_i - \hat{\beta}_1 - \hat{\beta}_2 x_i) &= 0 \\ \dfrac{\partial S}{\partial \beta_2} &= -2\sum_{i=1}^{n} x_i(y_i - \hat{\beta}_1 - \hat{\beta}_2 x_i) &= 0 \end{cases}$$

La première équation donne :

$$\hat{\beta}_1 n + \hat{\beta}_2 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i$$

d'où l'on déduit immédiatement :

$$\hat{\beta}_1 = \bar{Y} - \hat{\beta}_2 \bar{X} \tag{2.2}$$

où $\bar{X}$ et $\bar{Y}$ sont comme d'habitude les moyennes empiriques des x_i et des y_i. La seconde équation donne :

$$\hat{\beta}_1 \sum_{i=1}^{n} x_i + \hat{\beta}_2 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i$$

En remplaçant $\hat{\beta}_1$ par son expression (2.2) nous avons :

$$\hat{\beta}_2 = \frac{\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \bar{Y}}{\sum_{i=1}^{n} x_i^2 - \sum_{i=1}^{n} x_i \bar{X}} = \frac{\sum_{i=1}^{n} x_i (y_i - \bar{Y})}{\sum_{i=1}^{n} x_i (x_i - \bar{X})} = \frac{\sum_{i=1}^{n} (x_i - \bar{X})(y_i - \bar{Y})}{\sum_{i=1}^{n} (x_i - \bar{X})(x_i - \bar{X})} \tag{2.3}$$

Pour obtenir ce résultat, nous supposons qu'il existe au moins deux points d'abscisses différentes. Cette hypothèse notée H_1 s'écrit $x_i \neq x_j$ pour au moins deux individus. Elle permet d'obtenir l'unicité des coefficients estimés $\hat{\beta}_1$, $\hat{\beta}_2$ Une fois déterminés les estimateurs $\hat{\beta}_1$ et $\hat{\beta}_2$ nous pouvons estimer la droite de régression Par la formule

$$\hat{Y} = \hat{\beta}_1 + \hat{\beta}_2 X$$

Si nous évaluons la droite aux points x_i ayant servi à estimer les paramètres, nous obtenons des $\hat{y}_i$ et ces valeurs sont appelées les valeurs ajustées. Si nous

évaluons la droite en d'autres points, les valeurs obtenues seront appelées les valeurs prévues ou prévisions. Représentons les points initiaux et la droite de régression estimée. La droite de régression passe par le centre de gravité du Nuage de points $(\bar{X}, \bar{Y})$ *comme l'indique l'équation (2.2).*

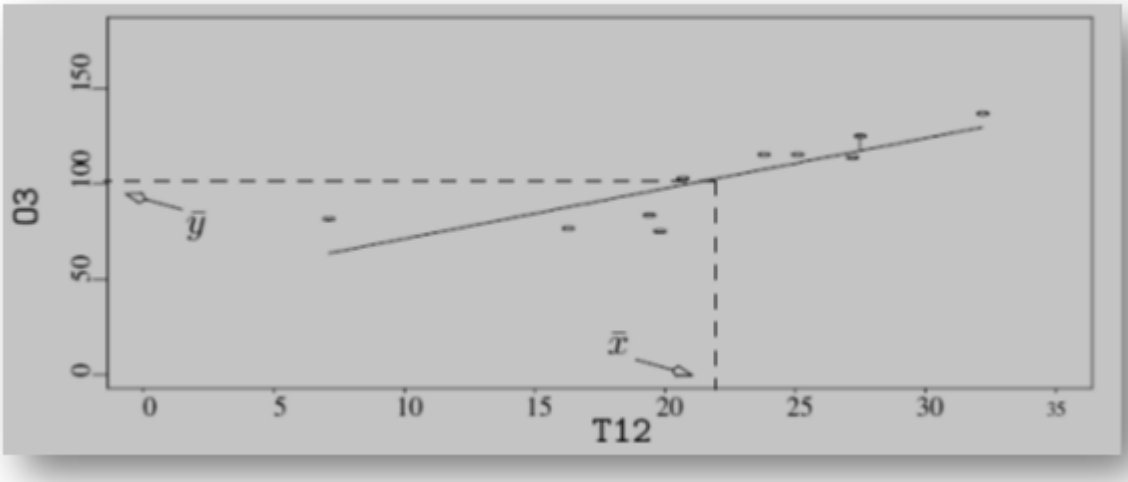


Figure 1. *Nuage de points, droite de régression et centre de gravité*

Théorème 1 *(Estimateurs sans biais)*
$\hat{\beta}_1$ et $\hat{\beta}_2$ sont des estimateurs sans biais de β_1 et β_2.

Preuve 2 *On a :*

$$\hat{\beta}_2 = \beta_2 + \frac{\sum_{i=1}^{n}(x_i - \bar{X})\varepsilon_i}{\sum_{i=1}^{n}(x_i - \bar{X})^2}$$

Dans cette expression, seuls les bruits ε_i sont aléatoires, et puisqu'ils sont centrés, on en déduit bien que $\mathbb{E}(\hat{\beta}_2) = \beta_2$.
Pour $\hat{\beta}_1$, on part de l'expression :

$$\hat{\beta}_1 = \bar{Y} - \hat{\beta}_2\bar{X}$$

d'où l'on tire :

$$\mathbb{E}(\hat{\beta}_1) = \mathbb{E}(\bar{Y}) - \bar{X}\mathbb{E}(\hat{\beta}_2) = \beta_1$$

Proposition 2 ***(Variances de $\hat{\beta}_1$ et $\hat{\beta}_2$)***

Les variances et covariance des estimateurs des paramètres valent :

$$\begin{aligned} V(\hat{\beta}_2) &= \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{X})^2} \\ V(\hat{\beta}_1) &= \frac{\sigma^2 \sum_{i=1}^{n} x_i^2}{n \sum_{i=1}^{n}(x_i - \bar{X})^2} \\ Cov(\hat{\beta}_1, \hat{\beta}_2) &= -\frac{\sigma^2 \bar{x}}{\sum_{i=1}^{n}(x_i - \bar{X})^2} \end{aligned}$$

Cette proposition nous permet d'envisager la précision de l'estimation en utilisant la variance. Plus la variance est faible, plus l'estimateur sera précis. Pour avoir des variances petites, il faut avoir un numérateur petit et (ou) un dénominateur Grand. Les estimateurs seront donc de faibles variances lorsque :

1. *La variance σ^2 est faible. Cela signifie que la variance de Y est faible et Donc les mesures sont proches de la droite à estimer ;*

2. *La quantité $\sum_{i=1}^{n}(x_i - \bar{X})^2$ est grande, les mesures x_i doivent être dispersées autour de leur moyenne ;*

3. *La quantité $\sum_{i=1}^{n} x_i^2$ ne doit pas être trop grande, les points doivent avoir Une faible moyenne en valeur absolue. En effet, nous avons :*
$$\frac{\sum_{i=1}^{n} x_i^2}{\sum_{i=1}^{n}(x_i - \bar{X})^2} = \frac{\sum_{i=1}^{n} x_i^2 - n\bar{X}^2 + n\bar{X}^2}{\sum_{i=1}^{n}(x_i - \bar{X})^2} = 1 + \frac{n\bar{X}^2}{\sum_{i=1}^{n}(x_i - \bar{X})^2}$$

L'équation (2.2) indique que la droite des MC passe par le centre de gravité du Nuage $(\bar{X}, \bar{Y})$. Supposons X positif, alors si nous augmentons la pente, l'ordonnée À l'origine va diminuer et vice versa. Nous retrouvons donc le signe négatif pour la covariance entre $\hat{\beta}_1$ et $\hat{\beta}_2$.

2.2.2.3 Résidus et variance résiduelle

Nous avons estimé β_1 et β_2. La variance σ^2 des ε_i est le dernier paramètre inconnu à estimer. Pour cela, nous allons utiliser les résidus : ce sont des estimateurs des erreurs inconnues ε_i.

Définition de résidus

Les résidus sont définis par

$$\hat{\varepsilon}_i = y_i - \hat{y}_i$$

Où $\hat{y}_i$ est la valeur ajustée de y_i par le modèle, c'est-à-dire $\hat{y}_i = \hat{\beta}_1 + \hat{\beta}_2 x_i$.
Nous avons la propriété suivante

Proposition 3 *Dans un modèle de régression linéaire simple, la somme des résidus est nulle.*
Intéressons-nous maintenant à l'estimation de σ^2 et construisons un estimateur Sans biais $\hat{\sigma}^2$.

Proposition 4 *(Estimateur de la variance du bruit)*
La statistique $\hat{\sigma}^2 = \sum_{i=1}^{n} \hat{\varepsilon}_i^2/(n-2)$ est un estimateur sans biais de σ^2.

2.2.2.4 Prévision

Un des buts de la régression est de proposer des prévisions pour la variable á expliquer Y. Soit x_{n+1} une nouvelle valeur de la variable X, nous voulons prédire y_{n+1}.
Le modèle indique que :

$$y_{n+1} = \beta_1 + \beta_2 x_{n+1} + \varepsilon_{n+1}$$

Avec $E(\varepsilon_{n+1}) = 0$, $V(\varepsilon_{n+1}) = \sigma^2$ et $Cov(\varepsilon_{n+1}, \varepsilon_i) = 0$ pour $i = 1, ..., n$.
Nous Pouvons prédire la valeur correspondante grâce au modèle estimé

$$\hat{y}_{n+1}^p = \hat{\beta}_1 + \hat{\beta}_2 x_{n+1}$$

Proposition 5 *(Variance de la prévision $\hat{y}_{n+1}^p$)*
La variance de la valeur prévue de $\hat{y}_{n+1}^p$ vaut

$$V(\hat{y}_{n+1}^p) = \sigma^2 \left(\frac{1}{n} + \frac{(x_{n+1} - \bar{X})^2}{\sum_{i=1}^{n} (x_i - \bar{X})^2} \right)$$

La variance de $\hat{y}_{n+1}^p$ nous donne une idée de la stabilité de l'estimation.
En prévision,on s'intéresse généralement à l'erreur que l'on commet entre la vraie valeur à prévoir y_{n+1} et celle que l'on prévoit $\hat{y}_{n+1}^p$. L'erreur peut être simplement résumée par la différence entre ces deux valeurs, c'est ce que nous appellerons l'erreur de prévision. Cette erreur de prévision Permet de quantifier la capacité du modèle à prévoir. Nous avons sur ce thème La proposition suivante

Proposition 6 *(Erreur de prévision)*
L'erreur de prévision, définie par $\hat{\varepsilon}^p_{n+1} = y_{n+1} - \hat{y}^p_{n+1}$ *satisfait les propriétés suivantes :*

$$\begin{aligned} \mathbb{E}(\hat{\varepsilon}^p_{n+1}) &= 0 \\ V(\hat{\varepsilon}^p_{n+1}) &= \sigma^2 \left(1 + \frac{1}{n} + \frac{(x_{n+1} - \bar{X})^2}{\sum (x_i - \bar{X})^2}\right) \end{aligned}$$

Remarque
La variance augmente lorsque x_{n+1} *s'éloigne du centre de gravité du nuage. Effectuer une prévision lorsque* x_{n+1} *est « loin » de* $\bar{X}$ *est donc périlleux, la variance de l'erreur de prévision peut alors être très grande !*

2.2.3 Interprétations géométriques

2.2.3.1 Représentation des individus

Pour chaque individu, ou observation, nous mesurons une valeur x_i et une valeur y_i. Une observation peut donc être représentée dans le plan, nous dirons alors que $\mathbb{R}^2$ est l'espace des observations.
$\hat{\beta}_1$ correspond à l'ordonnée à l'origine alors que $\hat{\beta}_2$ représente la pente de la droite ajustée. Cette droite minimise la somme des carrés des distances verticales des points du nuage à la droite ajustée.

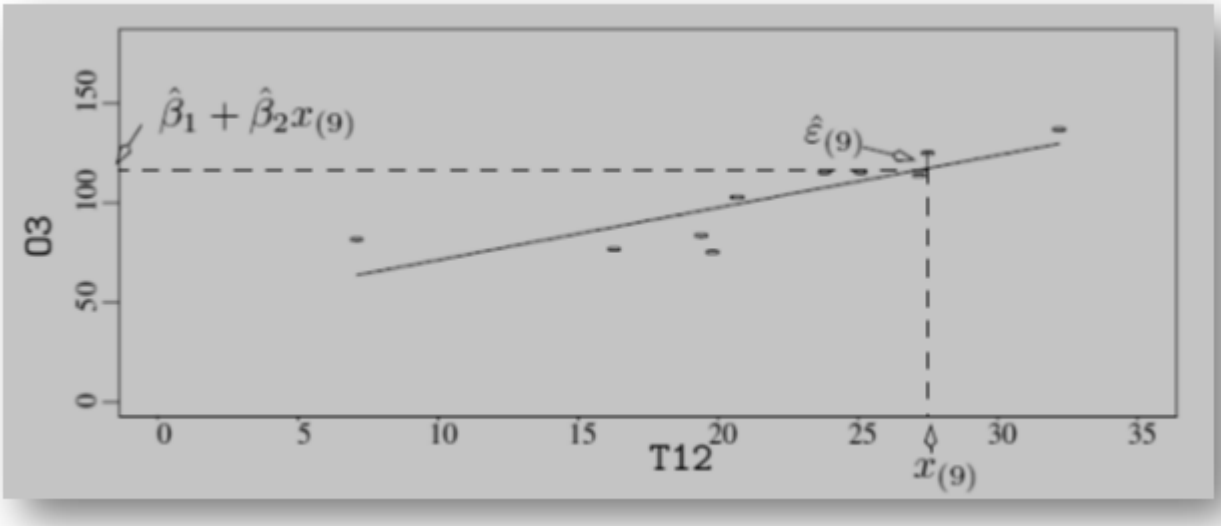


Figure 2. Représentation des individus.

Les couples d'observations (x_i, y_i) avec $i = 1, ..., n$ ordonnées suivant les valeurs croissantes de X sont notés (x_i, y_i). Nous avons représenté la neuvième valeur de X et sa valeur ajustée $\hat{y}_{(9)} = \hat{\beta}_1 + \hat{\beta}_2 x_{(9)}$ sur le graphique, ainsi que le résidu correspondant $\hat{\varepsilon}_{(9)}$.

2.2.3.2 Représentation des variables

Nous pouvons voir le problème d'une autre façon. Nous mesurons n couples de points (x_i, y_i). La variable X et la variable Y peuvent être considérées comme deux vecteurs possédant n coordonnées. Le vecteur X (respectivement Y) admet pour coordonnées : les observations $x_1, x_2, ..., x_n$ (respectivement $y_1, y_2, ..., y_n$). Ces deux vecteurs d'observations appartiennent au même espace $\mathbb{R}^n$: L'espace des variables.
Nous pouvons donc représenter les données dans l'espace des variables. Le vecteur $\mathbb{1}$est également un vecteur de $\mathbb{R}^n$ dont toutes les composantes valent 1. Les 2 vecteurs $\mathbb{1}$ et X engendrent un sous-espace de $\mathbb{R}^n$ de dimension 2. Nous avons supposé que $\mathbb{1}$ et X ne sont pas colinéaires grâce à H_1 mais ces vecteurs ne sont pas obligatoirement orthogonaux. Ces vecteurs sont orthogonaux lorsque $\bar{X}$, la moyenne des observations $x_1, x_2, ..., x_n$ vaut zéro.

La régression linéaire peut être vue comme la projection orthogonale du vecteur Y dans le sous-espace de $\mathbb{R}^n$ engendré par $\mathbb{1}$ et X, noté $\Im(X)$.Les coefficients $\hat{\beta}_1$ et $\hat{\beta}_2$ s'interprètent comme les composantes de la projection orthogonale notée $\hat{Y}$ de Y sur ce sous-espace. Voyons cela sur le graphique suivant :

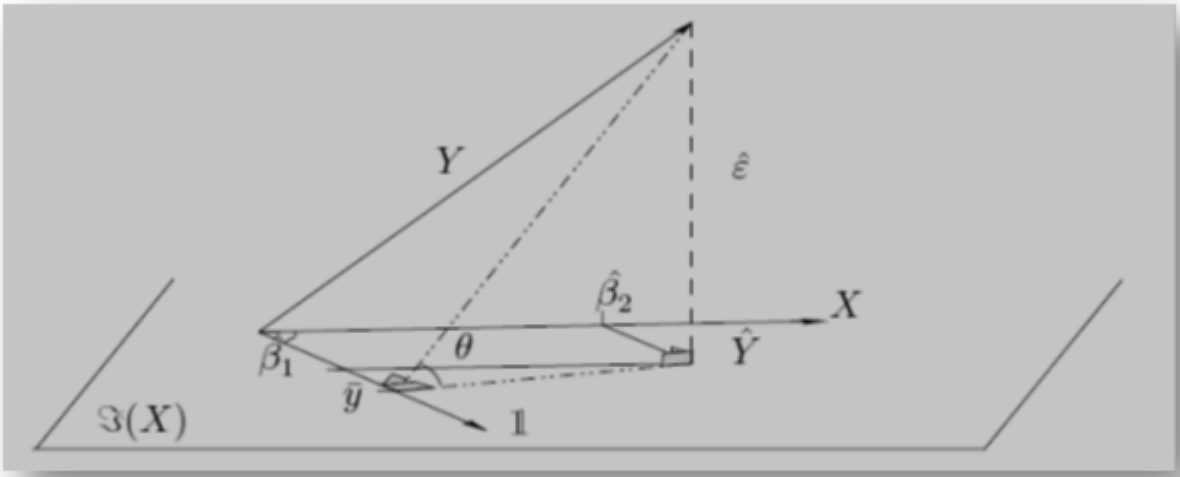


Figure 3. Représentation de la projection dans l'espace des variables.

Remarque
Les vecteurs $\mathbb{1}$ et X de normes respectives $\sqrt{n}$ et $\sqrt{\sum_{i=1}^{n} x_i^2}$ ne forment pas une base orthogonale. Afin de savoir si ces vecteurs sont orthogonaux, calculons leur produit scalaire. Le produit scalaire est la somme du produit terme à terme des composantes des deux vecteurs et vaut ici : $\sum_{i=1}^{n} x_i^2 1 = n\bar{X}$. Les vecteurs forment une base orthogonale lorsque la moyenne de X est nulle. En effet vaut alors zéro et le produit scalaire est nul. Les vecteurs n'étant en général pas orthogonaux, cela veut dire que $\hat{\beta}_1\mathbb{1}$ n'est pas la projection de Y sur la droite engendrée par $\mathbb{1}$ et $\hat{\beta}_2 X$ que n'est pas la projection de Y sur la droite engendrée par X.

2.2.3.3 Le coefficient de détermination $\mathbb{R}^2$

Un modèle, que l'on qualifiera de bon, possédera des estimations $\hat{y}_i$ proches des vraies valeurs y_i. Sur la représentation dans l'espace des variables (Figure 3).
La qualité peut être évaluée par l'angle θ. Cet angle est compris entre $-90°$ et $90°$. Un angle proche de $-90°$ ou de $90°$ indique un modèle de mauvaise qualité.
Le cosinus carré de θ est donc une mesure possible de la qualité du modèle et cette mesure varie entre 0 et 1. Le théorème de Pythagore nous donne directement que

$$\begin{aligned} \|Y - \bar{Y}\mathbb{1}\|^2 &= \|\hat{Y} - \bar{Y}\mathbb{1}\|^2 + \|\hat{\varepsilon}\|^2 \\ \sum_{i=1}^{n}(y_i - \bar{Y})^2 &= \sum_{i=1}^{n}(y_i - \bar{Y})^2 + \sum_{i=1}^{n}\hat{\varepsilon}_i^2 \\ SCT &= SCE + SCR, \end{aligned}$$

Où SCT (respectivement SCE et SCR) représente la somme des carrés totale (Respectivement expliquée par le modèle et résiduelle).
Le coefficient de détermination $\mathbb{R}^2$ est défini par

$$\mathbb{R}^2 = \frac{SCE}{SCT} = \frac{\|\hat{Y} - \bar{Y}\mathbb{1}\|^2}{\|Y - \bar{Y}\mathbb{1}\|^2}$$

C'est-à-dire la part de la variabilité expliquée par le modèle sur la variabilité totale. De nombreux logiciels multiplient cette valeur par 100 afin de donner Un pourcentage.
Remarques Dans ce cas précis, $\mathbb{R}^2$ est le carré du coefficient de corrélation empirique entre les x_i et les y_i et :

- le $\mathbb{R}^2$ correspond au cosinus carré de l'angle θ ;
- si $\mathbb{R}^2 = 1$, le modèle explique tout, l'angle θ vaut donc zéro, Y est dans $\Im(X)$ C'est-à-dire que $y_i = \beta_1 + \beta_2 x_i$;
- si $\mathbb{R}^2 = 0$, cela veut dire que $\sum_{i=1}^{n}(\hat{y}_i - \bar{Y})^2$ et donc que $\hat{y}_i = \bar{Y}$. Le modèle de régression linéaire est inadapté ;
- si $\mathbb{R}^2$ est proche de zéro, cela veut dire que Y est quasiment dans l'orthogonal de $\Im(X)$, le modèle de régression linéaire est inadapté, la variable X utilisée n'explique pas la variable Y.

2.2.4 Inférence statistique

Jusqu'à présent, nous avons pu, en choisissant une fonction de coût quadratique, ajuster un modèle de régression, à savoir calculer $\hat{\beta}_1$ et $\hat{\beta}_2$. Grâce aux coefficients estimés, nous pouvons donc prédire, pour chaque nouvelle valeur x_{n+1} une valeur de la variable à expliquer $\hat{y}^p_{n+1}$ qui est tout simplement le point sur la droite ajustée correspondant à l'abscisse x_{n+1}. En ajoutant l'hypothèse H_2, nous avons pu calculer l'espérance et la variance des estimateurs. Ces propriétés permettent d'appréhender de manière grossière la qualité des estimateurs proposés. Enfin ces deux hypothèses nous ont aussi permis de calculer l'espérance et la variance de la valeur prédite $\hat{y}^p_{n+1}$. Cependant nous souhaitons en général connaître la loi des estimateurs afin de calculer des intervalles ou des régions de confiance ou effectuer des tests. Il faut donc introduire une hypothèse supplémentaire concernant la loi des ε_i. L'hypothèse H_2 devient

$$H_3 \begin{cases} \varepsilon_i \sim N(0, \sigma^2) \\ \varepsilon_i \text{ sont indépendentes} \end{cases}$$

Où $N(0, \sigma^2)$ est une loi normale d'espérance nulle et de variance σ^2. Le modèle de régression devient le modèle paramétrique $\mathbb{R}^n, B_{\mathbb{R}^n}, N(\beta_1 + \beta_2 x, \sigma^2)$, où β_1, β_2, σ^2 sont à valeurs dans $\mathbb{R}$, $\mathbb{R}$ et $\mathbb{R}^+$ respectivement. La loi des ε_i étant connue, nous en déduisons la loi des y_i.
Nous allons envisager dans cette section les propriétés supplémentaires des estimateurs qui découlent de l'hypothèse H_3 (normalité et indépendance des erreurs) :

- Lois des estimateurs $\hat{\beta}_1$, $\hat{\beta}_2$ et σ^2
- Intervalles de confiance univariés et bivariés ;
- Loi des valeurs prévues $\hat{y}^p_{n+1}$ et intervalle de confiance.

$$\sigma^2_{\hat{\beta}_1} = \sigma^2 \left(\frac{\sum_{i=1}^n x_i^2}{n \sum_{i=1}^n (x_i - \bar{X})^2} \right), \ \hat{\sigma}^2_{\hat{\beta}_1} = \sigma^2 \left(\frac{\sum_{i=1}^n x_i^2}{n \sum_{i=1}^n (x_i - \bar{X})^2} \right)$$
$$\sigma^2_{\hat{\beta}_2} = \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{X})^2}, \ \hat{\sigma}^2_{\hat{\beta}_2} = \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{X})^2}$$

Où $\hat{\sigma}^2$ notons que les estimateurs de la colonne de gauche ne sont pas réellement des estimateurs. En effet puisque σ^2 est inconnu, ces estimateurs ne sont pas calculables avec les données. Cependant ce sont eux qui interviennent dans les lois des estimateurs $\hat{\beta}_1$ et $\hat{\beta}_2$ (cf. proposition ci-dessous).

Les estimateurs donnés dans la colonne de droite sont ceux qui sont utilisés (et utilisables) et ils consistent simplement à remplacer σ^2 par $\hat{\sigma}^2$ Les lois des estimateurs sont données dans la proposition suivante.

Proposition 7 *(Lois des estimateurs : variance connue)*
Les lois des estimateurs des MC sont :

i) $\hat{\beta}_1 \sim N(\beta_1, \sigma^2_{\hat{\beta}_1})$

ii) $\hat{\beta}_2 \sim N(\beta_2, \sigma^2_{\hat{\beta}_2})$

iii) $\hat{\beta} = \begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix} \sim N(\beta, \sigma^2 V), \beta = \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix}$ *et* $V = \frac{1}{\sum_{i=1}^n (x_i - \bar{X})^2} \begin{bmatrix} \sum_{i=1}^n x_i^2/n & -\bar{X} \\ -\bar{X} & 1 \end{bmatrix}$

iv) $\frac{n-2}{\sigma^2}\hat{\sigma}^2$ *Suit une loi du* χ^2 *à* $(n-2)$ *degrés de liberté (ddl)* $(\chi^2_{(n-2)})$

v) $(\hat{\beta}_1, \hat{\beta}_2)$ *et* $\hat{\sigma}^2$ *sont indépendants*

La variance σ^2 *n'est pas connue en général, nous l'estimons par* $\hat{\sigma}^2$ *. Les estimateurs Des MC ont alors les propriétés suivantes.*

Proposition 8 *(Lois des estimateurs : variance estimée)*
Lorsque σ^2 *est estimée par* $\hat{\sigma}^2$ *, nous avons :*

i) $\frac{\hat{\beta}_1 - \beta_1}{\hat{\sigma}_{\hat{\beta}_1}} \sim T_{n-2}$ *Où* T_{n-2} *est une loi de Student à* $(n-2)$ *ddl.*

ii) $\frac{\hat{\beta}_2 - \beta_2}{\hat{\sigma}_{\hat{\beta}_2}} \sim T_{n-2}$

iii) $\frac{1}{2\hat{\sigma}^2}(\hat{\beta} - \beta)' V^{-1} (\hat{\beta} - \beta) \sim F_{2,n-2}$ *où* $F_{2,n-2}$ *est une loi de Fisher à 2 ddl au numérateur et* $(n-2)$ *ddl au dénominateur.*

Proposition 9 *(Intervalle de confiance(IC) et region de confiance(RC) de niveau* $1-\alpha$ *pour les paramètres)*
Un IC de $\beta_i (i \in 1, 2)$ *est donné par :*

$$\left[\hat{\beta}_i - t_{n-2}(1 - \frac{\alpha}{2})\hat{\sigma}_{\hat{\beta}_i}, \hat{\beta}_i + t_{n-2}(1 - \frac{\alpha}{2})\hat{\sigma}_{\hat{\beta}_i}\right]$$

où $t_{n-2}(1-\alpha)$ *représente le fractile de niveau* $(1-\alpha/2)$ *d'une loi du* χ^2 *à* $(n-2)$ *degrés de liberté.*

Proposition 10 *(IC pour $E(y_i)$)*
Un IC de $E(y_i) = \beta_1 + \beta_2 x_i$ est donné par :

$$\left[\hat{y}_j \pm t_{n-2}(1-\alpha/2)\hat{\sigma}\sqrt{\frac{1}{n} + \frac{(x_j - \bar{X})^2}{\sum_{i=1}^{n}(x_j - \bar{X})^2}}\right]$$

En calculant les IC pour tous les points de la droite, nous obtenons une hyperbole de confiance. En effet, lorsque x_j est proche de $\bar{X}$ le terme dominant de la variance est $1/n$, mais dès que x_j s'éloigne de $\bar{X}$ le terme dominant est le terme au carré.

Proposition 11 *(IC pour y_{n+1})*
Un IC de y_{n+1} est donné par :

$$\left[\hat{y}^p_{n+1} \pm t_{n-2}(1-\alpha/2)\hat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{(x_j - \bar{X})^2}{\sum_{i=1}^{n}(x_j - \bar{X})^2}}\right]$$

Cette formule exprime que plus le point à prévoir est éloigné de $\bar{X}$ plus la variance de la prévision et donc l'IC seront grands. Une approche intuitive consiste à remarquer que plus une observation est éloignée du centre de gravité, moins nous avons d'information sur elle. Lorsque la valeur à prévoir est à l'intérieur de l'étendue des x_i, le terme dominant de la variance est la valeur 1 et donc la variance est relativement constante. Lorsque x_{n+1} est en dehors de l'étendue des x_i, le terme dominant peut être le terme au carré, et la forme de l'intervalle sera à nouveau une hyperbole.

2.2.5 Estimateurs du maximum de vraisemblance

Lorsque nous supposons que les résidus suivent une loi normale, le modèle de régression devient le modèle paramétrique $\mathbb{R}^n, B_{\mathbb{R}^n}, N(\beta_1 + \beta_2 x, \sigma^2)$, où β_1, β_2, σ^2 sont à valeurs dans $\mathbb{R}, \mathbb{R}$ et $\mathbb{R}^+$ respectivement. La loi des ε_i étant connue, nous en déduisons la loi des y_i. Nous calculons la vraisemblance de l'échantillon ainsi que les estimateurs qui maximisent cette vraisemblance. Puisque les y_i valent par hypothèse $\beta_1 + \beta_2 x_i + \varepsilon_i$, nous savons grâce à H_3 que la loi des y_i est une loi normale de moyenne $\beta_1 + \beta_2 x_i$ et de variance σ^2. l'indépendance des ε_i entraîne l'indépendance des y_i. La vraisemblance vaut

alors :

$$\begin{aligned} L(\beta_1, \beta_2, \sigma^2) &= \prod_{i=1}^{n} f(y_i) \\ &= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \beta_1 - \beta_2 x_i)^2\right] \\ &= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n exp\left[-\frac{1}{2\sigma^2}S(\beta_1, \beta_2)\right] \end{aligned}$$

En passant au logarithme, nous obtenons :

$$logL(\beta_1, \beta_2, \sigma^2) = -\frac{n}{2}log2\pi\sigma^2 - \frac{1}{2\sigma^2}S(\beta_1, \beta_2)$$

Calculons les dérivées par rapport à, β_1, β_2 et σ^2

$$\begin{cases} \dfrac{\partial logL(\beta_1, \beta_2, \sigma^2)}{\partial \beta_1} = -\dfrac{1}{2\sigma^2}\dfrac{\partial S(\beta_1, \beta_2)}{\partial \beta_1} = \dfrac{1}{\sigma^2}\sum_{i=1}^{n}(y_i - \beta_1 - \beta_2 x_i) \\ \dfrac{\partial logL(\beta_1, \beta_2, \sigma^2)}{\partial \beta_2} = -\dfrac{1}{2\sigma^2}\dfrac{\partial S(\beta_1, \beta_2)}{\partial \beta_2} = \dfrac{1}{\sigma^2}\sum_{i=1}^{n} x_i(y_i - \beta_1 - \beta_2 x_i) \\ \dfrac{\partial logL(\beta_1, \beta_2, \sigma^2)}{\partial \sigma^2} = -\dfrac{n}{2\sigma^2} + \dfrac{1}{2\sigma^4}\sum_{i=1}^{n}(y_i - \beta_1 - \beta_2 x_i)^2 \end{cases}$$

Les estimateurs du maximum de vraisemblance de β_1, β_2 sont identiques aux estimateurs obtenus par les MC. L'estimateur de σ^2 vaut

$$\hat{\sigma}^2_{mv} = \frac{1}{n}\sum_{i=1}^{n}\hat{\varepsilon}_i^2$$

L'estimateur du MV de σ^2 est donc biaisé car différent de l'estimateur des MC qui lui est non biaisé. Cela veut dire que $\mathbb{E}(\hat{\sigma}^2_{mv}) \neq \sigma^2$

2.3 La régression linéaire multiple

2.3.1 Modélisation

Le modèle de régression multiple est une généralisation du modèle de régression Simple lorsque les variables explicatives sont en nombre fini. Nous Supposons donc que les données collectées suivent le modèle suivant :

$$y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_p x_{ip} + \varepsilon_i, \quad i = 1, ..., n \tag{2.4}$$

Où :
– les x_{ij} sont des nombres connus, non aléatoires. La variable x_{i1} peut valoir 1 pour tout i variant de 1 à n. Dans ce cas, β_1 représente la constante.
En statistiques, cette colonne de 1 est presque toujours présente :
– les paramètres à estimer β_j du modèle sont inconnus.
– les ε_i sont des variables aléatoires inconnues.
En utilisant l'écriture matricielle de (2.4), nous obtenons la définition suivante

2.3.2 Modèle de régression linéaire multiple

Un modèle de régression linéaire est défini par une équation de la forme

$$Y_{n\times 1} = X_{n\times p}\beta_{p\times 1} + \varepsilon_{n\times 1} \tag{2.5}$$

Où :
• Y est un vecteur aléatoire de dimension n,
• X est une matrice de taille $n\times p$ connue, appelée matrice du plan d'expérience, X est la concaténation des p variables X_j : $X = (X_1|X_2|...|X_p)$. Nous noterons la i^e ligne du tableau X par le vecteur ligne $x_i^{'} = (x_{i1}, ..., x_{ip})$.
• β est le vecteur de dimension p des paramètres inconnus du modèle ;
• ε est le vecteur centré, de dimension n, des erreurs.
Nous supposons que la matrice X est de plein rang. Cette hypothèse sera notée H_1. Comme, en général, le nombre d'individus n est plus grand que le nombre de variables explicatives p, le rang de la matrice X vaut p. La présentation précédente revient à supposer que la fonction liant Y aux variables explicatives X est un hyperplan représenté (Figure 4)

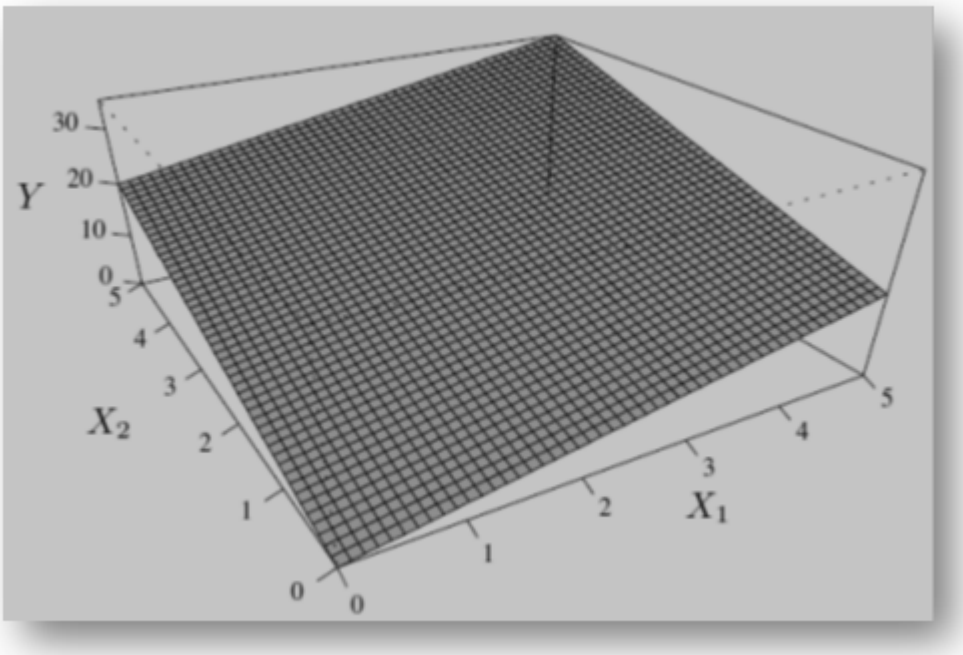


Figure 4.

2.3.3 Estimateurs des moindres carrés

Définition 2 *On appelle estimateur des moindres carrés (noté MC) $\hat{\beta}$ de β la valeur suivante*

$$\hat{\beta} = \underset{\beta_1, ..., \beta_p}{argmin} \quad \sum_{i=1}^{n} \left(y_i - \sum_{i=1}^{n} \beta_j x_{ij}\right)^2$$

$$= \underset{\beta \in \mathbb{R}^p}{argmin} \quad (Y - X\beta)'(Y - X\beta)$$

Théorème 2 *Si l'hypothèse H_1 est vérifiée, l'estimateur des MC $\hat{\beta}$ de β vaut*

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

2.3.3.1 Calcul de $\hat{\beta}$

Il est intéressant de considérer les variables dans l'espace des variables $(\mathbb{R}^n)$.
Ainsi, Y vecteur colonne, définit dans $\mathbb{R}^n$ un vecteur $\overrightarrow{OY}$ d'origine O et d'extrémité Y. Ce vecteur a pour coordonnées $(y_1, ..., y_n)$. La matrice X du plan d'expérience est formée de p vecteurs colonnes. Chaque vecteur X_j définit dans $\mathbb{R}^n$ un vecteur $\overrightarrow{OX_{ij}}$ d'origine O et d'extrémité X_j. Ce vecteur a pour coordonnées $(x_{1j}, ..., x_{nj})$. Ces p vecteurs linéairement indépendants (hypothèse H_1) engendrent un sous-espace vectoriel de $\mathbb{R}^n$, noté dorénavant $\Im$, de dimension p.

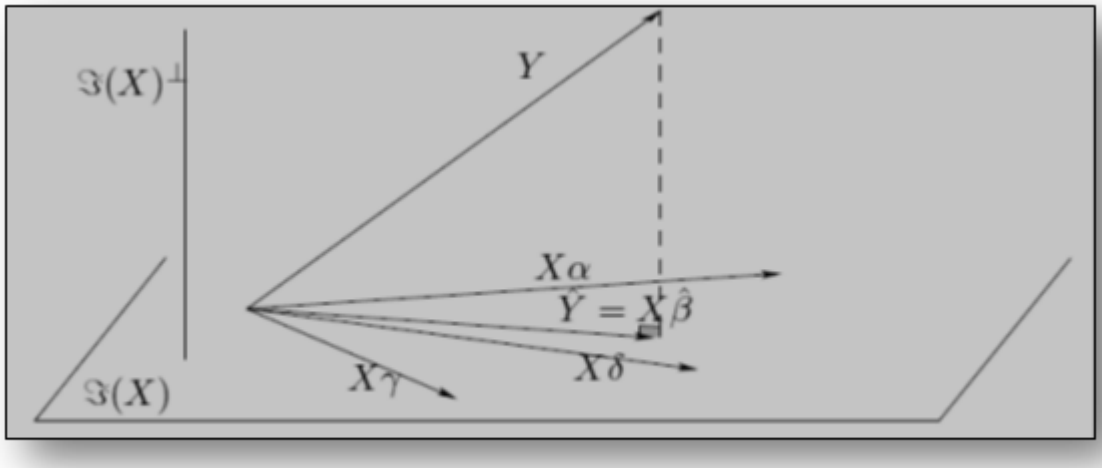


Figure 5. Représentation dans l'espace des variables.

Cet espace $\Im$ appelé image de X est engendré par les colonnes de X.
Il est parfois appelé espace des solutions. Ainsi, tout vecteur $\overrightarrow{v}$ de $\Im$ s'écrit de façon unique sous la forme suivante :

$$\overrightarrow{v} = \alpha_1 \overrightarrow{X}_1 + ... + \alpha_p \overrightarrow{X}_p$$

Le vecteur Y est la somme d'un élément de $\Im(X)$ et d'un bruit, élément de $\mathbb{R}^n$, qui n'a aucune raison d'appartenir à $\Im(X)$. Minimiser $S(\beta)$ revient à

chercher un élément de $\Im(X)$ qui soit le plus proche de Y, au sens de la norme euclidienne classique. Par définition, cet unique élément est appelé projection orthogonale de Y sur $\Im(X)$. Il sera noté $\hat{Y} = P_X Y$, où P_X est la matrice de projection orthogonale sur $\Im(X)$. Dans la littérature anglo-saxonne, cette matrice est souvent notée H et est appelée « hat matrix » car elle met des « hat » sur Y. Par souci de cohérence de l'écriture, nous noterons l'élément courant (i, j) de P_X, h_{ij}. L'élément $\hat{Y}$ de $\Im(X)$ est aussi noté $\hat{Y} = X\hat{\beta}$ où $\hat{\beta}$ est l'estimateur des MC de β. L'espace orthogonal à $\Im$ noté $\Im(X)^{\perp}$ est souvent appelé espace des résidus. Le vecteur $\hat{Y} = P_X Y$ contient les valeurs ajustées par le modèle de Y.

• Calcul de $\hat{\beta}$ par projection :

Trois possibilités de calcul de $\hat{\beta}$ sont proposées.

– La première consiste à connaître la forme analytique de P_X. La matrice de projection orthogonale sur $\Im(X)$ est donnée par :

$$P_X = X(X'X)^{-1}X'$$

Et, comme $P_X Y = X\hat{\beta}$ nous obtenons $\hat{\beta} = (X'X)^{-1}X'Y$

– La deuxième méthode utilise le fait que le vecteur Y de $\mathbb{R}^n$ se décompose de façon unique en une partie sur $\Im(X)$ et une partie sur $\Im(X)^{\perp}$, cela s'écrit :

$$Y = P_X Y + (I - P_X)Y$$

La quantité $(I - P_X)Y$ étant un élément de $\Im(X)^{\perp}$ est orthogonale à tout élément v quelconque $\Im(X)$. Rappelons que $\Im(X)$ est l'espace engendré par les colonnes de X, c'est-à-dire que toutes les combinaisons linéaires de variables $X_1, ..., X_p$ sont éléments de $\Im(X)$ ou encore que, pour tout $\alpha \in \mathbb{R}^p$, nous avons $X\alpha \in \Im(X)$. Les deux vecteurs v et $(I-P_X)Y$ étant orthogonaux, le produit scalaire entre ces deux quantités est nul, soit :

$$\begin{array}{rcll}
\langle v, (I - P_X)Y\rangle & = & 0 & , \forall v \in \Im(X) \\
\langle X\alpha, (I - P_X)Y\rangle & = & 0 & , \forall \alpha \in \mathbb{R}^p \\
\alpha' X'(I - P_X)Y & = & 0 & \\
X'Y' & = & X'P_X Y & \text{avec } P_X Y = X\hat{\beta} \\
X'Y' & = & X'X\hat{\beta} & X \text{ de rang plein} \\
\hat{\beta} & = & (X'X)^{-1}X'Y &
\end{array}$$

Nous retrouvons $P_X = X(X'X)^{-1}X'$ matrice de projection orthogonale sur l'espace engendré par les colonnes de X. Les propriétés caractéristiques d'un

projecteur orthogonal $P_X^{'} = P_X$ et $P_X^2 = P_X$ sont vérifiées.
– La dernière façon de procéder consiste à écrire que le vecteur $(I - P_X)Y$ est orthogonal à chacune des colonnes de X qui engendre $\Im(X)$.

$$\begin{cases} \langle X_1, Y - X\hat{\beta}\rangle &= 0 \\ \quad . & \\ \quad . & \Leftrightarrow \quad X^{'}Y = X^{'}X\hat{\beta} \\ \quad . & \\ \langle X_p, Y - X\hat{\beta}\rangle &= 0 \end{cases}$$

Soit $P_X = X(X^{'}X)^{-1}X^{'}$ la matrice de projection orthogonale sur $\Im(X)$, la matrice de projection orthogonale sur $\Im(X)^{\perp}$ est $P_{X^{\perp}} = (I - P_X)$.
• Calcul matriciel
Nous pouvons aussi retrouver le résultat précédent de manière analytique en écrivant la fonction à minimiser $S(\beta)$:

$$\begin{aligned} S(\beta) &= Y^{'}Y + \beta^{'}X^{'}X\beta - Y^{'}X\beta - \beta^{'}X^{'}Y \\ &= \quad Y^{'}Y + \beta^{'}X^{'}X\beta - Y^{'}X\beta \end{aligned}$$

Une condition nécessaire d'optimum est que la dérivée première par rapport à β s'annule. Ou la dérivée s'écrit comme suit :

$$\frac{\partial S(\beta)}{\partial \beta} = -2X^{'}Y + 2X^{'}X\beta$$

D'où, s'il existe, l'optimum, noté $\hat{\beta}$, vérifie :

$$-2X^{'}Y + 2X^{'}\hat{\beta} = 0$$

C'est-à-dire :

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

Pour s'assurer que ce point $\hat{\beta}$ est bien un minimum strict, il faut que la dérivée seconde soit une matrice définie positive. Or la dérivée seconde s'écrit :

$$\frac{\partial^2 S(\beta)}{\partial \beta^2} = 2X^{'}X$$

Et X est de plein rang donc $X^{'}X$ est inversible et n'a pas de valeur propre nulle. La matrice $X^{'}X$ est donc définie. De plus $\forall z \in \mathbb{R}^P$, nous avons :

$$2X_z^{'}X_z = 2\langle X_z, X_z\rangle = 2\|X_z\|^2 \geq 0$$

$(X^{'}X)$ est donc bien définie positive et $\hat{\beta}$ est bien un minimum strict.

2.3.4 Interprétation

Nous venons de voir que $\hat{Y}$ est la projection de Y sur le sous-espace engendré par les colonnes de X. Cette projection existe et est unique même si l'hypothèse H_1 n'est pas vérifiée. L'hypothèse H_1 nous permet d'obtenir un $\hat{\beta}$ unique. Dans ce cas, s'intéresser aux coordonnées de $\hat{\beta}$ a un sens, et ces coordonnées sont les coordonnées de $\hat{Y}$ dans le repère $X_1, ..., X_p$. Ce repère n'a aucune raison d'être orthogonal et donc $\hat{\beta}_j$ n'est pas la coordonnée de la projection de Y sur X_j. Nous avons :

$$P_X Y = \hat{\beta}_1 X_1 + ... + \hat{\beta}_p X_p$$

Calculons la projection de Y sur X_j.

$$\begin{aligned} P_{X_j} Y &= P_{X_j} P_X Y \\ &= \hat{\beta}_1 P_{X_j} X_1 + ... + \hat{\beta}_p P_{X_j} X_p \\ &= \hat{\beta}_j X_j + \sum_{i \neq j} \hat{\beta}_i P_{X_j} X_i \end{aligned}$$

Cette dernière quantité est différente de $\hat{\beta}_j X_j$ sauf si X_j est orthogonal à toutes les autres variables.
Lorsque toutes les variables sont orthogonales deux à deux, il est clair que $(X^{'} X)$ est une matrice diagonale :

$$(X^{'} X) = diag(\|X_1\|^2, ..., \|X_p\|^2)$$

2.3.5 Quelques propriétés statistiques

Le statisticien cherche à vérifier que les estimateurs des MC que nous avons construits admettent de bonnes propriétés au sens statistique. Dans notre cadre de travail, cela peut se résumer en deux parties : l'estimateur des MC est-il sans biais et est-il de variance minimale dans sa classe d'estimateurs ? pour cela, nous supposons une seconde hypothèse notée H_2 indiquant que les erreurs sont centrées, de même variance et non corrélées entre elles. L'écriture de cette hypothèse est $H_2 : \mathbb{E}(\varepsilon) = 0, \Sigma_\varepsilon = \sigma^2 I_n$, avec I_n la matrice identité d'ordre n. Cette hypothèse nous permet de calculer

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

L'estimateur des MC est donc sans biais. Calculons sa variance

$$V(\hat{\beta}) = V((X'X)^{-1}X'Y) = (X'X)^{-1}X'V(Y)X(X'X)^{-1} = \sigma^2(X'X)^{-1}$$

Proposition 12 *($\hat{\beta}$ sans biais)*
L'estimateur $\hat{\beta}$ des MC est un estimateur sans biais de β et sa variance vaut :

$$V(\hat{\beta}) = \sigma^2(X'X)^{-1}$$

2.3.6 Résidus et variance résiduelle

Les résidus sont définis par la relation suivante :

$$\hat{\varepsilon} = Y - \hat{Y}$$

En nous servant du modèle, $Y = X\beta + \varepsilon$ et du fait que $X\beta \in \Im(X)$ nous avons une autre écriture des résidus :

$$\hat{\varepsilon} = Y - X\beta = Y - X(X'X)^{-1}X'Y = (I - P_X)Y = P_{X^\perp}Y = P_{X^\perp}\varepsilon$$

Les résidus appartiennent donc à $\Im(X)^\perp$ et cet espace est aussi appelé espace des résidus. Les résidus sont donc toujours orthogonaux à $\hat{Y}$ Nous avons les propriétés suivantes :

Proposition 13 *(Propriétés de $\hat{\varepsilon}$ et $\hat{Y}$)*
Sous les hypothèses H_1 et H_2, nous avons :

$$\begin{aligned}
\mathbb{E}(\hat{\varepsilon}) &= P_{X^\perp}\mathbb{E}(\varepsilon) = 0 \\
V(\hat{\varepsilon}) &= \sigma^2 P_{X^\perp} I P'_{X^\perp} = \sigma^2 P_{X^\perp} \\
\mathbb{E}(\hat{Y}) &= X\mathbb{E}(\hat{\beta}) = X\beta \\
V(\hat{Y}) &= \sigma^2 P_X \\
Cov(\hat{\varepsilon}, \hat{Y}) &= 0
\end{aligned}$$

Proposition 14 *($\hat{\sigma}^2$ sans biais)*
La statistique $\hat{\sigma}^2$ est un estimateur sans biais de σ^2.
A partir de cet estimateur de la variance résiduelle, nous obtenons immédiatement un estimateur de la variance de $\hat{\beta}$ en remplaçant σ^2 par son estimateur :

$$\hat{\sigma}^2_{\hat{\beta}} = \hat{\sigma}^2(X'X)^{-1} = \frac{SCR}{n-p}(X'X)^{-1}$$

Nous avons donc un estimateur de l'écart-type de l'estimateur $\hat{\beta}$ de chaque coefficient de la régression β_j

$$\hat{\sigma}_{\hat{\beta}_j} = \sqrt{\hat{\sigma}^2[(X'X)^{-1}]_{jj}}$$

2.3.7 Interprétation géométrique

Le théorème de Pythagore donne directement l'égalité suivante :

$$\begin{aligned} \|Y\|^2 &= \|\hat{Y}\|^2 + \|\hat{\varepsilon}\|^2 \\ &= \|X\hat{\beta}\|^2 + \|Y - X\hat{\beta}\|^2 \end{aligned}$$

Si la constante fait partie du modèle, alors nous avons toujours par le théorème de Pythagore :

$$\|Y - \bar{Y}\mathbb{1}\|^2 = \|Y - \bar{Y}\mathbb{1}\|^2 + \|\hat{\varepsilon}\|^2$$

SCT(la somme des carrés totale) = SCE(la somme des carrés expliquée par le modèle)+SCR(la somme des carrés résiduelle).

Définition 3 *(Définition de $\mathbb{R}^2$)*
Le coefficient de détermination (multiple) $\mathbb{R}^2$ est défini par :

$$\mathbb{R}^2 = \frac{\|\hat{Y}\|^2}{\|Y\|^2} = \cos^2\theta_0$$

Et si la constante fait partie de $\Im(X)$ par :

$$\mathbb{R}^2 = \frac{V \text{ expliquée par le modéle}}{\text{variation totale}} = \frac{\|\hat{Y} - \bar{Y}\mathbb{1}\|^2}{\|Y - \bar{Y}\mathbb{1}\|^2} = \cos^2\theta$$

Le $\mathbb{R}^2$ peut aussi s'écrire en fonction des résidus :

$$\mathbb{R}^2 = 1 - \frac{\|\hat{\varepsilon}\|^2}{\|Y - \bar{Y}\mathbb{1}\|^2}$$

ce coefficient mesure le cosinus carré de l'angle entre les vecteurs Y et $\hat{Y}$ pris à l'origine ou pris en $\bar{Y}$ Ce dernier est toujours plus grand que le premier, le $\mathbb{R}^2$ calculé lorsque la constante fait partie de $\Im(X)$ est donc plus petit que le $\mathbb{R}^2$ calculé directement

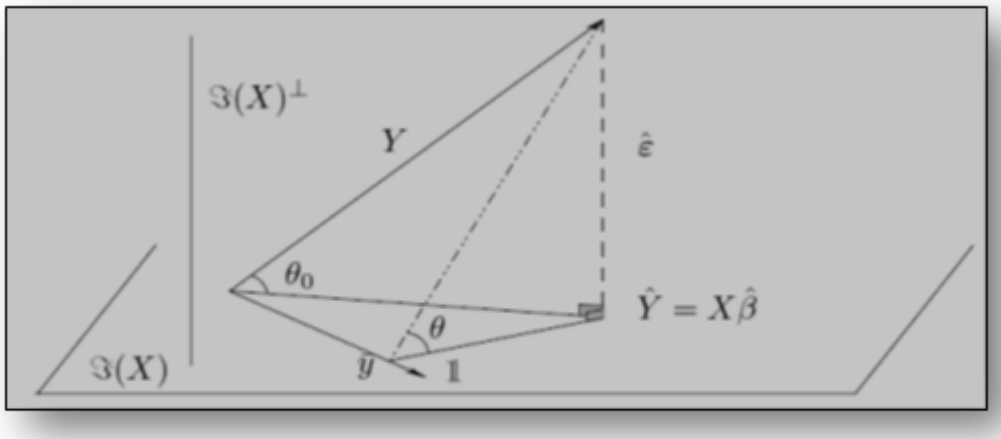


Figure 6. *Représentation des variables et interprétation géométrique du* $\mathbb{R}^2$

Définition 4 *(*$\mathbb{R}^2$ *ajusté)*
Le coefficient de détermination ajusté $\mathbb{R}^2_a$ *est défini par :*

$$\mathbb{R}^2_a = 1 - \frac{n}{n-p}\frac{\|\hat{\varepsilon}\|^2}{\|Y\|^2}$$

Et, si la constante fait partie de $\Im(X)$ *par :*

$$\mathbb{R}^2_a = 1 - \frac{n-1}{n-p}\frac{\|\hat{\varepsilon}\|^2}{\|Y - \bar{Y}\mathbb{1}\|^2}$$

2.3.8 Inférence statistique

Nous rappelons le contexte :

$$Y_{n\times 1} = X_{n\times p}\beta_{n\times 1} + \varepsilon_{n\times 1}$$

Sous les hypothèses :
- $H_1 : rang(X) = p$
- $H_2 : \mathbb{E}(\varepsilon) = 0,\ \Sigma_\varepsilon = \sigma^2 I_n$

Nous allons désormais supposer que les erreurs suivent une loi normale et donc H_2 devient :

- $H_3 : \varepsilon \sim N(0, \sigma^2 I_n)$

Nous pouvons remarquer que H_3 contient H_2. De plus, dans le cas gaussien, $Cov(\varepsilon_i, \varepsilon_j = \sigma^2 \delta_{ij})$ implique que les ε_i sont indépendants. L'hypothèse H_3 s'écrit $\varepsilon_1, ..., \varepsilon_n$ sont i.i.d. et de loi $N(0, \sigma^2)$.
L'hypothèse gaussienne va nous permettre de calculer la vraisemblance et donc les estimateurs du maximum de vraisemblance (EMV).

2.3.9 Estimateurs du maximum de vraisemblance

Calculons la vraisemblance de l'échantillon. La vraisemblance est la densité de l'échantillon vu comme fonction des paramètres. Grâce à l'indépendance des erreurs, les observations sont indépendantes et la vraisemblance s'écrit :

$$L(Y, \beta, \sigma^2) = \prod_{i=1}^{n} f_Y(y_i) = \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{n}{2}} exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \sum_{j=1}^{p}\beta_j x_{ij})^2\right]$$

Nous avons donc :

$$L(Y, \beta, \sigma^2) = \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{n}{2}} exp\left[-\frac{1}{2\sigma^2}\|Y - X\beta\|^2\right]$$

Ce qui donne :

$$\log L(Y, \beta, \sigma^2) = -\frac{n}{2}\log\sigma^2 - \frac{n}{2}\log 2\pi - \frac{1}{2\sigma^2}\|Y - X\beta\|^2$$

Nous obtenons :

$$\frac{\partial L(Y, \beta, \sigma^2)}{\partial \beta} = \frac{1}{2\sigma^2}\frac{\partial}{\partial \beta}(\|Y - X\beta\|^2), \tag{2.6}$$

$$\frac{\partial L(Y, \beta, \sigma^2)}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\|Y - X\beta\|^2 \tag{2.7}$$

A partir de (2.6), nous avons évidemment $\hat{\beta}_{MV} = \hat{\beta}$ et à partir de (2.7) nous avons :

$$\hat{\sigma}^2_{MV} = \frac{\|Y - X\hat{\beta}_{MV}\|^2}{n}$$

Et donc que $\hat{\sigma}^2_{MV} = (n-p)\hat{\sigma}^2/n$ l'estimateur du MV est donc biaisé par opposition à $\hat{\sigma}^2$ l'estimateur obtenu par les MC. Afin de vérifier que nous avons bien un maximum, il faut étudier les dérivées secondes sous l'hypothèse supplémentaire H_3.

Proposition 15 *(Lois des estimateurs : variance connue)*
Sous les hypothèses H_1 et H_3, nous avons :

i) $\hat{\beta}$ *est un vecteur gaussien de moyenne* β *et de variance* $\sigma^2(X'X)^{-1}$.

ii) $(n-p)\hat{\sigma}^2/\sigma^2$ *suit un* χ^2 *à* $(n-p)$ *ddl* (χ^2_{n-p}).

iii) $\hat{\beta}$ *et* $\hat{\sigma}^2$ *sont indépendants.*

Proposition 16 *(Lois des estimateurs : variance estimée)*
Sous les hypothèses H_1 et H_3, nous avons :

i) pour $i = 1, ..., p$

$$T_j = \frac{\hat{\beta}_j - \beta_j}{\hat{\sigma}\sqrt{[(X'X)^{-1}]_{jj}}} \sim T_{(n-p)}$$

ii) Soit R *une matrice de taille* qp *de rang* $q(q \leq p)$ *alors la v.a :*

$$\frac{1}{q\hat{\sigma}^2}(R(\hat{\beta}-\beta))' \left[R(X'X)^{-1}R'\right]^{-1} R(\hat{\beta}-\beta) \sim F_{q,n-p}$$

Théorème 3 *(Intervalle de confiance (IC) et region de confiance (RC) des paramètres)*

i) Un IC, de niveau $1-\alpha$, *pour un* β_j *pour* $j = 1, ..., p$ *est donné par :*

$$\left[\hat{\beta}_j - t_{n-p}(1-\alpha/2)\hat{\sigma}\sqrt{[(X'X)^{-1}]_{jj}}, \hat{\beta}_j + t_{n-p}(1-\alpha/2)\hat{\sigma}\sqrt{[(X'X)^{-1}]_{jj}}\right]$$

ii) Un IC, de niveau $1-\alpha$, *pour* σ^2 *est donné par :*

$$\left[\frac{(n-p)\hat{\sigma}^2}{c_2}, \frac{(n-p)\hat{\sigma}^2}{c_1}\right] \quad \text{où } P(c_1 \leq \chi^2_{n-p} \leq c_2) = 1-\alpha$$

iii) Une RC pour $q(q \leq p)$ *paramètres* β_j *notés* $(\beta_{j1}, ..., \beta_{jq})$ *de niveau* $1-\alpha$ *est donnée,*

- *lorsque σ est connue, par :*

$$RC_\alpha(R\beta) = \left\{ R\beta \in \mathbb{R}^q, \frac{1}{\hat{\sigma}^2}(R(\hat{\beta}-\beta))' \left[R(X'X)^{-1}R'\right]^{-1} R(\hat{\beta}-\beta) \leq \chi^2_q(1-\alpha) \right\}$$

- *lorsque σ est inconnue, par :*

$$RC_\alpha(R\beta) = \left\{ R\beta \in \mathbb{R}^q, \frac{1}{q\hat{\sigma}^2}(R(\hat{\beta}-\beta))' \left[R(X'X)^{-1}R'\right]^{-1} R(\hat{\beta}-\beta) \leq F_{q,n-p}(1-\alpha) \right\}$$

Où R est la matrice de taille qp dont tous les éléments sont nuls sauf les $[R]_{iji}$ qui valent 1. Les valeurs c_1 et c_2 sont les fractiles d'un χ^2_q et $F_{q,n-p}(1-\alpha)$ est le fractile de niveau $(1-\alpha)$ d'une loi de Fisher admettant $(q, n-p)$ ddl.

Chapitre 3

Application

Introduction

La démarche expérimentale est indispensable pour analyser l'existence de la causalité. C'est ce qu'on va utiliser lors d'une étude pour prouver une efficacité d'une thérapeutique particulière. C'est cette démarche expérimentale qui a toute son importance en médecine.
La démarche statistique est indispensable pour quantifier la part du hasard dans les résultats obtenus. Elle permet de mettre en évidence des différences mais elle ne porte aucun jugement de causalité.
Les statistiques font progresser la médecine. Deux sortes de progrès existent en médecine :
Les découvertes scientifiques, qui nécessitent de l'observation et de la créativité.
Les démonstrations « scientifiques », les justifications. C'est là que les statistiques prennent toute leur place. Il faut exclure toute idée préconçue et se placer dans une démarche expérimentale, afin de tester cette hypothèse de la manière la plus rigoureuse possible.

3.1 Régression linéaire simple sur un essai clinique

Le tableau suivant contient des données réelles d'une étude croisée comparant un nouveau laxatif par rapport à un autre standard sur 35 Personne :

Patient no	new treatment (y-variables)(days with stool)	bisacodyl (x-variables) (days of stool)
1	24	8
2	30	13
3	25	15
4	35	10
5	39	9
6	30	10
7	27	8
8	14	5
9	39	13
10	42	15
11	41	11
12	38	11
13	39	12
14	37	10
15	47	18
16	30	13
17	36	12
18	12	4
19	26	10
20	20	8
21	43	16
22	31	15
23	40	14
24	31	7
25	36	12
26	21	6
27	44	19
28	11	5
29	27	8
30	24	9
31	40	15
32	32	7
33	10	6
34	37	14
35	19	7

3.1.1 Avec estimation des paramètres β_1 et β_2 méthode des moindres carrés :

$$Y = \beta_1 + \beta_2 X$$

Il faut donc trouver les coefficients β_1 et β_2 nous savons que :

$$\beta_2 = \frac{cov(X,Y)}{var(X)} = \frac{\frac{1}{N}\sum_{i=1}^{n} x_i y_i - \bar{X}\bar{Y}}{\frac{1}{N}\sum_{i=1}^{N} x_i^2 - \bar{X}^2} = 2,064968517$$

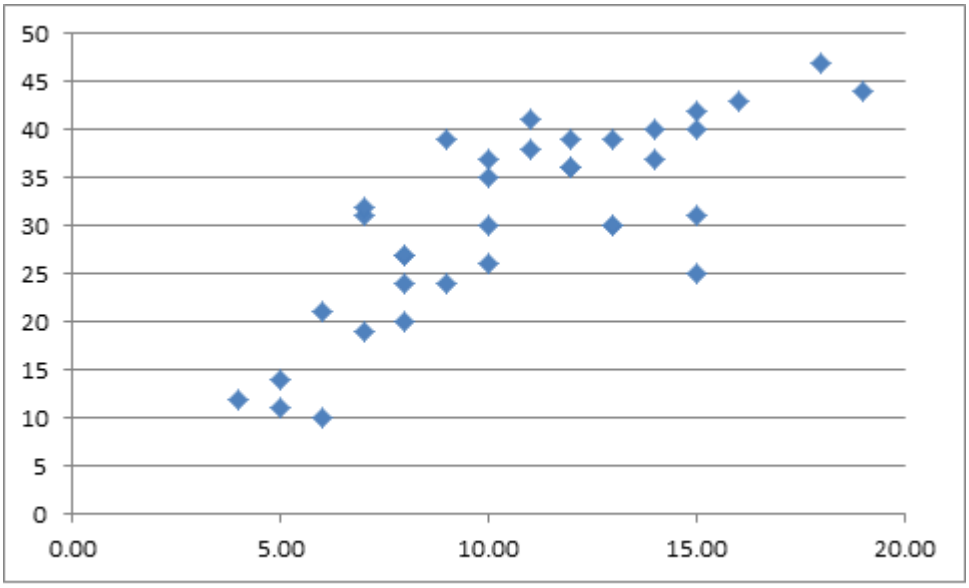
$$\beta_1 = \bar{Y} + \beta_2 \bar{X} = 8,6468$$

Alors la fonction est :

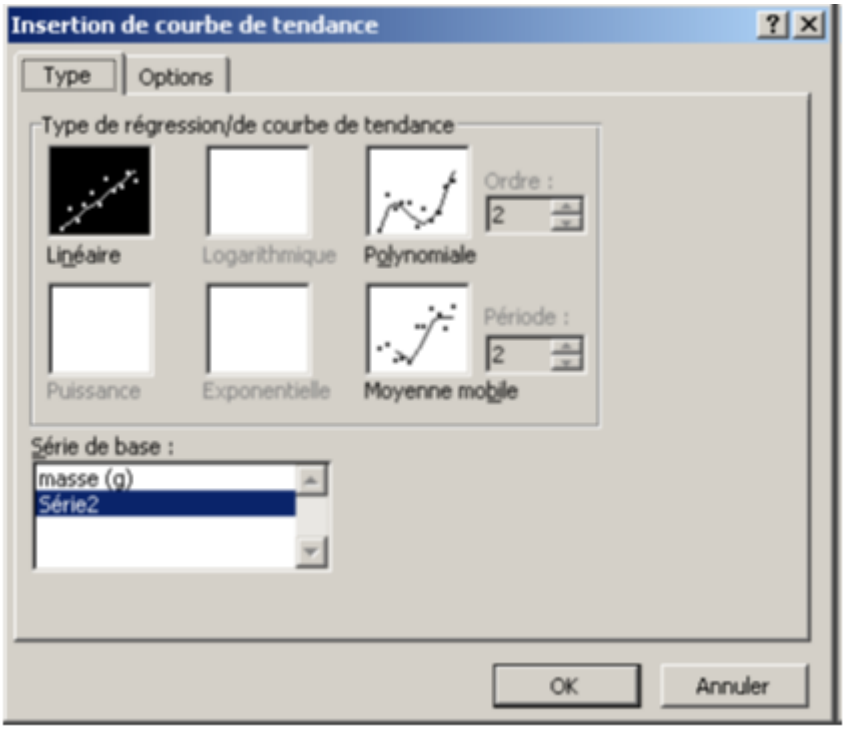
$$Y = 8,6468 + 2,064968517X$$

3.1.2 Avec EXCEL :

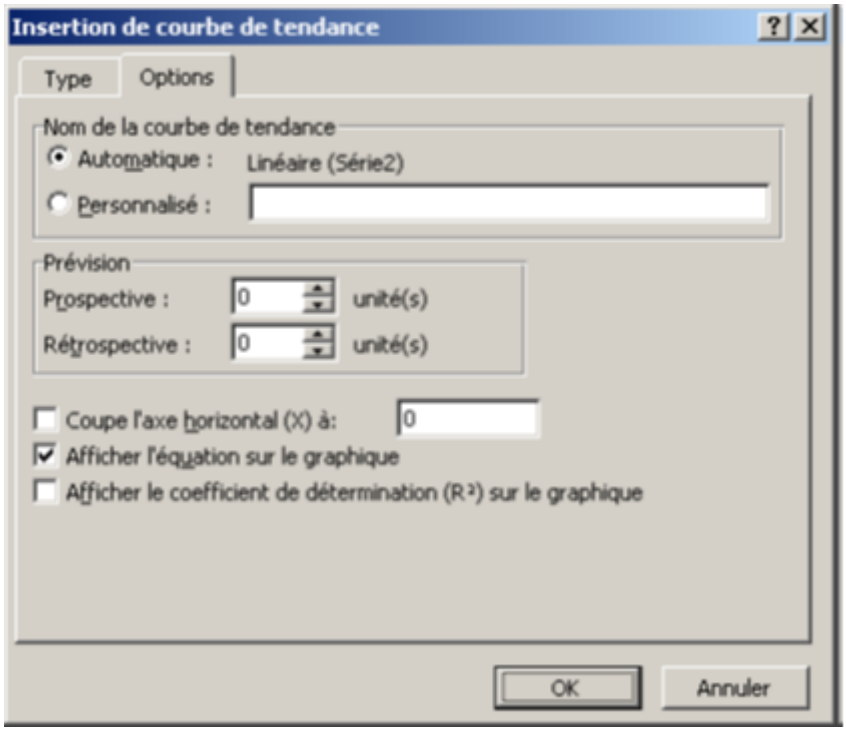
1- Nous avons les x_i et les y_i de tableau précédent, on clique sur insertion puis sur nuage de points et on obtenu le graphe suivant :



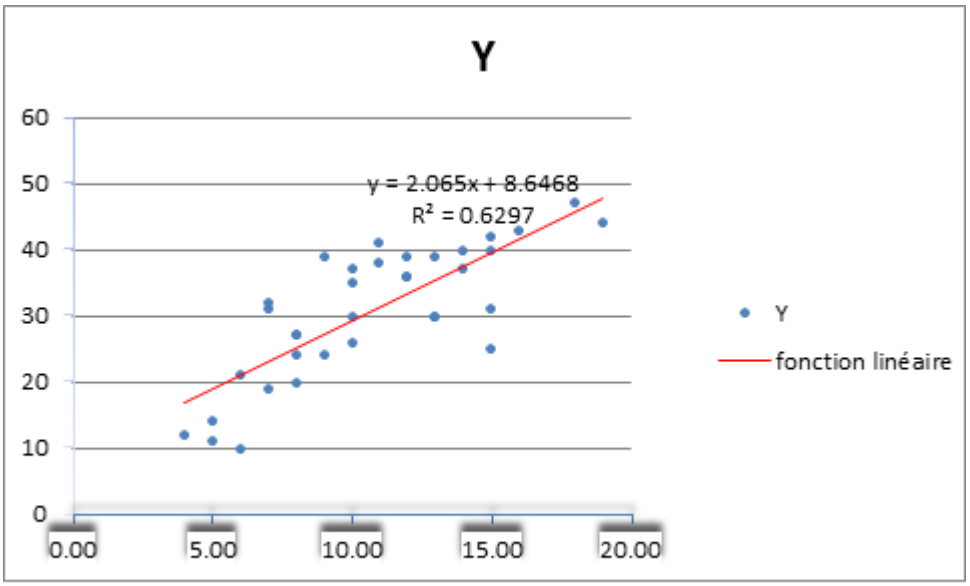
2-On clique sur les points pour les sélectionner.
3-On clique sur le bouton droit et on choisit ajouter une courbe de tendance.



4-Nous choisissons linéaire et en suit options, et nous cochons la case afficher l'équation.



5-Le graphe à cet aspect est :



On voit que :

$$Y = 2,065X + 8,646$$

3.1.3 Avec WinBUGS :

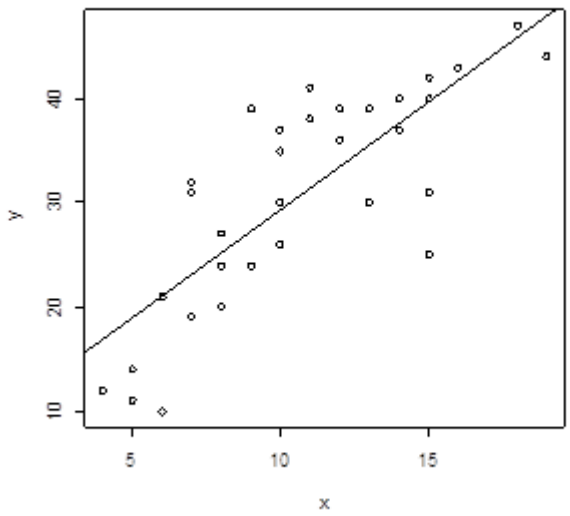
On entre les données :
list $(X = c(8, 13, 15, 10, 9, 10, 8, 5, 13, 15, 11, 11, 12, 10, 18, 13, 12, 4, 10, 8, 16, 15, 14, 7,$
$12, 6, 19, 5, 8, 9, 15, 7, 6, 14, 7), y = c(24, 30, 25, 35, 39, 30, 27, 14, 39, 42, 41, 38, 39, 37,$
$47, 30, 36, 12, 26, 20, 43, 31, 40, 31, 36, 21, 44, 11, 27, 24, 40, 32, 10, 37, 19))$
list$(alpha = 0, beta = 0, tau = 1)$
list$(alpha = 0.5, beta = 0.5, tau = 0.5)$

Le modèle est écrit :
{
for(i in 1 : 35) {
$y[i] \sim$ dnorm$(mu[i], tau)$
mu$[i] <-$ alpha + beta * $X[i]$
}
alpha$\sim dnorm(0, 1.0E - 6)$
beta $\sim dnorm(0, 1.0E - 6)$
tau $\sim dgamma(1.0E - 3, 1.0E - 3)$
sigma < -1 /sqrt(tau)
}

Les statistiques :

Node	mean	sd	MC error	2.5%	median	97.5%	start	sample
alpha	8.656	3.235	0.02396	2.305	8.634	15.04	1	20000
beta	2.064	0.2853	0.00211	1.505	2.062	2.624	1	20000
tau	0.02639	0.006558	5.1E-5	0.01527	0.02581	0.04082	1	20000

Pour afficher le graphe on utilise :
Call :lm(formula $= Y \sim X$,data = fr)



La droite de la régression est :

$$Y = 8.647 + 2.065X$$

3.1.4 Interprétation des résultats :

On remarque qu'il ya des logiciels qui peuvent faire la régression linéaire mais les résultats reste les même, dans notre exemple la fonction est :

$$Y = 2.065X + 8.646$$

La pente est 2.065 et l'ordonné à l'origine est 8.646, et si nous avons

$$X = 1 \rightarrow Y = 10$$

c'est-à-dire :
Le nouveau traitement est meilleur que le traitement standard.

3.2 Régression linéaire multiple sur un essai clinique avec WinBUGS :

Nous avons le tableau précédent on ajoutant un autre facteur qui est l'âge de patient .
On obtient le tableau suivant :

Patient no	y-variables	x_1-variables	x_2-variables
1	24	8	25
2	30	13	30
3	25	15	25
4	35	10	31
5	39	9	36
6	30	10	33
7	27	8	22
8	14	5	18
9	39	13	14
10	42	15	30
11	41	11	36
12	38	11	30
13	39	12	27
14	37	10	38
15	47	18	40
16	30	13	31
17	36	12	25
18	12	4	24
19	26	10	27
20	20	8	20
21	43	16	35
22	31	15	29
23	40	14	32
24	31	7	30
25	36	12	40
26	21	6	31
27	44	19	41
28	11	5	26
29	27	8	24
30	24	9	30
31	40	15	20
32	32	7	31
33	10	6	29
34	37	14	43
35	19	7	30

Modèle avec le logiciel WinBugs :
Y-variables : nouveau treatment.
X_1-variables : bisacodyl.
X_2-variables : age.
$Y \sim N(mu; tau)$
$mu = \alpha + \beta_1 X_1 + \beta_2 X_2$
modéle
for(i in $1:35$)
$y[i] \sim dnorm(mu[i], tau)$
$mu[i] < -alpha + beta1 * X1[i] + beta2 * X2[i]$

$alpha \sim dnorm(0, 1.0E-6)$
$beta1 \sim dnorm(0, 1.0E-6)$
$beta2 \sim dnorm(0, 1.0E-6)$
$tau \sim dgamma(1.0E-3, 1.0E-3)$
$sigma < -1/sqrt(tau)$

Nous procédons ensuite à estimer, cette fois sur deux canaux, avec 110000 itérations (1000) suffisamment chacun, en gardant une itération de 150. Les paramètres de la ligne est estimé, $\alpha = 2,332$ avec un écart type de $4,985$ et $\beta_1 = 1.876$ avec un écart type de $0,3003$, $\beta_2 = 0,282$ avec un écart type de $0,171$. Les sorties de WinBUGS sont :

Node	mean	sd	MC error	2.5%	median	97.5%	start	sample
alpha	2.332	4.985	0.03595	-7.53	2.353	12.14	1	22000
beta1	1.876	0.3003	0.002056	1.282	1.875	2.472	1	22000
beta2	0.2827	0.1719	0.001191	-0.05832	0.2827	0.6232	1	22000
tau	0.02786	0.006994	5.042E-5	0.01588	0.02728	0.04326	1	22000

3.3 Interprétation des résultats

* Dans le modèle de régression linéaire la droite de régression est $Y = 8.646 + 2.065X$. La pente est 2.065 et dirigé l'original est $8,646$, et si nous avons $X = 1 \rightarrow Y = 10$ donc le nouveau traitement est meilleur que le traitement standard.
* La droite de régression dans le modèle de régression linéaire multiple est $Y = 2.332 + 0.282X_1 + 1.876X_2$

Nous ajoutons le facteur l'âge ou non le nouveau traitement est la meilleure.

Conclusion Générale

C'est vrai que la mathèmatiques la mère des sciences, puisque dans ce travaille on a trouvé des réponses pour les essais cliniques qui est un domaine plus proche à la médecine avec une méthode d'analyse multivariée c'est la régression linéaire.
Quotidienne et dans la littérature théorique. Pour une approche claire et synthétique nous avons abordé de maniére détaillée les concepts et les techniques de base de la régression linéaire et nous avons précisé tout particulièrement les conditions et les situations de la méthode et nous avons mis l'accent sur les aspects pratiques puisque on trouve des résultats ou avec on produit médicaments médicales et il faut être efficace pour les patients.

Bibliographie

[1] Arnaud Guyader. Régression linéaire,2013.

[2] Société canadienne du cancer. Les essais cliniques, 2014.

[3] Leem les entreprises du médicament. Les etudes cliniques en 20 questions, 2011.

[4] Société canadienne du cancer. Les essais cliniques, guide à l'intentiondes personnes atteintedu cancer, 2007.

[5] ierre-André Cornillon, Éric Matzner-Lober, Régression : Théorie et applications, 2007.

[6] Yuan M. & Lin Y. (2005). Efficient empirical bayes variable selection and estimation in linear models. Journal of the American Statistical Association, 100, 1215–1225.

[7] Miller A. (2002). Subset selection in regression. Chapmann & Hall/CRC, London, 2 ed.

[8] Labdaoui Ahlam. Thése de doctorat l'analyse bayesienne dans les essais cliniques, 2015.

[9] PATERSON (R.) et RUSSELL (M.H.). Clinical Trials in malignant Disease. Part III. Breast Cancer Evaluation of post-operative Radiotherapy. J. Fac. Radiol., 1559, 10, 175.

[10] SCHWARTZ D, Lazar Pn, PapozL, statistique médicale et biologique, Flammarion, (ISBN 2257104463).

[11] Akaike H. (1973). Information theory and an extension of the maximum likelihood principle. Dans Second international symposium on information theory, réd. B.N. Petrov & B.F. Csaki, pp. 267–281. Academiai Kiado, Budapest.

[12] SCHWARTZ (D). Méthodes statistiques à l'usage des médecins et des biologistes. Ed. Méd. Flammarion, Paris, 1963.

[13] Birkes D. & Dodge Y. (1993). Alternative Methods of Regression. Wiley.

[14] José LABARERE, Corrélation et régression linéaire simple, 2012.

[15] S. Le Digabel, école Polytechnique de Montréal. Régression linéaire simple.

[16] Thierry Foucart. Colinearité et regréssion linéair, 2006.

[17] Josiane Confais, Monique Le Guen. Premiers pas en regression linéaire avec sas. 2006

[18] Ricco Rakotomalala.Régression linéaire simple, prédire/expliquer les valeurs d'une variable quantitative Y à partir d'une autre variable X.

[19] Y.Dodge, V.Rousson, « Analyse de régression appliquée », Dunod, 2004.

[20] Frédéric Bertrand. Régression linéaire simple, 2010.

Sites web :
http ://www.e-cancer.fr
http ://ansm.sante.fr
http ://www.sanofi.com

Résumé

Ce travail comporte trois chapitres :
Le premier chapitre sur les généralités des essais cliniques, le deuxième sur le principe de régression linéaire simple et multiple et le troisième sur l'application avec plusieurs méthodes.
Mots clés : essais clinique, régression linéaire.

Abstract
This work has three chapters :
The first chapter on the general clinical trials, the second on the principle of simple and multiple linear regression and the third on the application with several methods.
Keywords : clinical trials, linear regression.

ملخص
يتكون هذا العمل من ثلاثة فصول :
الفصل الأول في التجارب السريرية العامة، والثاني على مبدأ الانحدار الخطي البسيط والمتعدد والثالث على التطبيق مع العديد من الطرق.
الكلمات المفتاحة : التجارب السريرية، الانحدار الخطي.