Introduction au modèle linéaire

Didier Concordet
Unité de Biostatistique
Ecole Vétérinaire de Toulouse

Sommaire

1 INTRODUCTION **4**

2 REGRESSION LINEAIRE SIMPLE **11**
2.1 Hypothèses 11
2.2 Estimation des paramètres 13
2.3 Test sur les paramètres 16
2.4 Intervalles de confiance 17
2.5 Vérification des hypothèses 20
2.6 Détection et élimination des valeurs aberrantes 24
2.7 Un exemple 26
2.8 Non respect des hypothèses 32

3 REGRESSION MULTIPLE **38**
3.1 Hypothèses 39
3.2 Estimation des paramètres 39
3.3 Tests d'hypothèses 40
3.4 Diagnostic de multicolinéarité 45
3.5 Ce qu'il ne faudrait pas croire 46
3.6 Un exemple 47

4 ANALYSE DE VARIANCE **51**
4.1 Généralités 51
4.2 Les modèles 54

5 ANALYSE DE LA VARIANCE A UN FACTEUR **55**
5.1 Modèle à effets fixes 55
5.2 Modèle à effets aléatoires 60

5.3 Plan et analyse des expériences factorielles 64
5.4 Exemple d'analyse de variance d'un plan factoriel à deux facteurs à effets fixes 66
5.5 Analyse d'un plan factoriel non équilibré 75
5.6 Analyse d'un modèle à effets mixtes 78
5.7 Une généralisation 82
5.8 Plans hiérarchiques à deux facteurs 85
5.9 Plans en mesures répétées 90

Chapitre 1

INTRODUCTION

La modélisation est une technique qui permet de chercher, de quantifier des relations. [1]

Cette technique repose sur l'écriture de modèles, qui doivent décrire la forme des relations existant entre les variables. Il existe différents types de modèles, l'objet de ce poly n'étant pas de faire une revue exhaustive des modèles, nous nous contenterons de définir les deux types suivants:
1) les modèles deterministes
2) les modèles aléatoires également dénommés stochastiques.

Modèles déterministes vs modèles statistiques.

Dans les modèles déterministes, aucune place n'est laissée à l'approximation, la relation qui lie les différentes variables d'intérêts est fixée. Prenons un exemple simple pour fixer les idées. Supposons que l'on recherche la relation entre une différence de potentiel que nous noterons U la resistance R d'un objet traversé par un courant d'intensité I. Une façon simple de procéder est de faire varier l'intensité I et de voir comment se comporte la différence de potentiel U. Si on représentait graphiquement U en fonction de I, on ob-

[1] Une attention doit être apportée sur le sens donnée ici au mot relation, notamment en ce qui concerne une relation de cause à effet. Aucun calcul statistique ne peut établir une relation cause-effet. Cette relation ne peut être formulée que par un individu qui en assume l'entière responsabilité à partir d'une théorie qui lui est personnelle. La statistique ne peut intervenir que pour éprouver cette théorie en la confrontant à des résultats expérimentaux. On doit disposer de mesures suffisantes. Il est parfaitement inutile d'élaborer un modèle complexe si on ne peut le "nourrir" qu'avec des données insuffisantes ou de qualité douteuse.existant entre des variables.

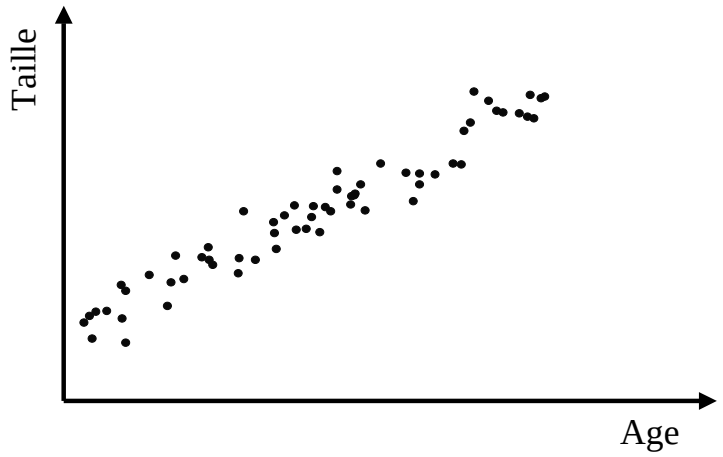

Figure 1.1: Pour un âge fixé, un grand nombre de tailles sont possibles.

tiendrait une droite. Si on recommençait la même opération avec un autre appareil de même résistance, on obtiendrait la même droite. En fait, quelque soit l'appareil de résistance R utilisé pour l'expérience, la relation entre U et I serait toujours la même. Nous venons de redécouvrir la loi d'Ohm: $U = RI$. Quand R et I sont fixés, U est fixé: on qualifiera la relation entre U et I de déterministe. En biologie, ou plus généralement dans les sciences de la vie, de telles relations n'*existent* pas. Prenons un exemple pour le constater.
On veut étudier la relation entre age et taille chez les enfants âgés d'au moins 3 ans et d'au plus 15 ans. Supposons que de telles mesures aient été effectuées sur 30 enfants. Le graphique 1 représente les résultats obtenus. Que constate t-on? Pour un age fixé,il existe un grand nombre de tailles possibles, on voit donc que même s'il existe une relation entre age et taille, elle n'est pas de la même nature que celle qui lie U et I dans l'exemple précédent. Aussi faudra t-il rajouter dans l'équation un terme supplémentaire qui sera un terme d'erreur (en général noté ε), et qui traduira une connaissance imparfaite de la relation. En fait on mettra dans ce terme, toute les fluctuations de la taille non expliquées par le modèle que nous utilisons. Ce terme contiendra les effets de facteurs (sûrement liés à la taille de façon déterministe), mais qui

de part leur trop grand nombre sont impossibles à incorporer dans le modèle ou simplement à identifier. Ceci nous amène donc à conclure qu'un modèle n'est jamais juste ou faux, il est simplement plus ou moins bien adapté à la question posée.

Modélisation

Ecrire un modèle revient donc à écrire une relation entre plusieurs variables. Nous nous intéresserons dans ce poly au relation de la forme:

$$Y = f(x_1, x_2, ..., x_p) + \varepsilon \tag{1.1}$$

Dans cette relation, les variables $x_1, x_2, ..., x_p$ jouent des rôles symétriques, on les supposera déterministes, fixées ou encore connues avec une grande précision.
Ces variables (à droite du signe =) sont en général appelées variables **indépendantes** (l'emploi d'anglicismes est parfois malheureux) ou plus simplement, variables **explicatives**. La variable Y est souvent appelée variable **dépendante** ou variable **à expliquer**; elle est supposée sujette à des variations aléatoires comme en atteste la présence du terme ε à droite du signe =. La désignation même des variables précise le rôle que nous allons leur faire jouer. En fait, la formule (1.1) est incomplète. En effet si on disposait d'une formule de ce type, aucune expérience ne serait nécessaire: Y serait toujours connue à ε près. Ce qui manque dans (1.1), ce sont les **paramètres** qui vont quantifier la façon dont les variables $x_1, x_2, ..., x_p$ sont liées à Y. Reprenons l'exemple de la relation entre taille et age. Le graphique 1 suggère une relation du type:

$$Y = \theta_0 + \theta_1 x + \varepsilon \tag{1.2}$$

Ici, Y représente la taille de tous les enfants âgés d'au moins 3 ans et d'au plus 15 ans, x représente l'age de ces enfants, θ_0 et θ_1 sont des **paramètres inconnus** du modèle. L'équation (1.2) ne fait que préciser la forme de la relation entre la taille et l'age, mais elle ne précise pas de valeurs pour θ_0 et θ_1.
Si ces paramètres pouvaient être connus avec certitude, nous connaîtrions précisément la relation entre taille et age. Pour connaître ces paramètres

avec certitude, il faudrait disposer de l'age et de la taille de tous les enfants sur lesquels porte l'étude, ce qui en général est impossible, on ne dispose que d'échantillons. Les seules valeurs auxquelles accessibles, sont donc des estimations des vraies valeurs (nous détaillerons un peu par la suite les conséquences de cette imparfaite connaissance des paramètres. La façon dont interviennent les paramètres inconnus dans l'écriture du modèle a une importance cruciale.
Un modèle dans lequel les paramètres inconnus (notés $\theta_1, \theta_2, \ldots, \theta_p$) interviennent linéairement s'écrit sous la forme :

$$Y = \theta_1 g_1(x_1, \ldots, \theta_q) + \theta_2 g_2(x_1, \ldots, \theta_q) + \ldots + \theta_p g_p(x_1, \ldots, \theta_q)\varepsilon \tag{1.3}$$

Si les **paramètres** interviennent linéairement dans le modèle, on parlera de **modèle linéaire**. Dans le cas contraire, on parlera de **modèle non linéaire**.
L'intérêt des modèles linéaires repose sur le fait que les paramètres sont facilement estimés sans biais, et que sous certaines conditions sur ε, qui seront détaillées un peu plus bas, il est facile de faire des tests. Prenons des exemples:

(1) $$Y = \theta_0 + \theta_1 x + \varepsilon$$

(2) $$Y = \theta_0 \theta_1 x + \varepsilon$$

(3) $$Y = \theta_0 + \theta_1 x + \theta_2 x^2 + \varepsilon$$

(4) $$Y = \theta_0 e^{-x} + \theta_1 e^{-2x} + \varepsilon$$

(5) $$Y = \theta_0 e^{-\theta_1 x} + \varepsilon.$$

Les modèles 1, 3, 4 sont linéaires (en fonction des paramètres) bien que les modèles 3 et 4 soient non linéaires en fonction des variables explicatives. Les modèles 2 et 5 sont non linéaires (en fonction des paramètres). Dans ce chapitre, nous ne nous intéresserons qu'aux modèles linéaires (en fonction des paramètres).

Les effets aléatoires

Dans le modèle (1.1) toutes les variables explicatives sont supposées déterministes. La plupart du temps, cette situation simple ne permet pas de répondre de façon effective aux questions posées. Pour illustrer cette insuffisance, prenons un exemple : Supposons que l'on cherche à comparer l'efficacité de 2 médicaments : la guéritoumycine et la panacémycine. Quatre élevages ont été sélectionnés pour participer à cet essai. Dans chaque élevage un échantillon d'animaux a été tiré au hasard, une moitié des animaux de l'échantillon ont reçus la guéritoumycine et l'autre moitié la panacémycine. Les données brutes ont été analysées et les analyses ont montré que la guéritoumycine a une plus grande efficacité que la panacémycine. Que peut-on conclure ? Pour répondre convenablement à cette question, il est nécessaire de préciser si la nature du facteur élevages :
- si les élevages ont été choisis, le facteur élevage est un facteur **fixe** et les résultats de l'analyse ne peuvent pas être extrapolés à d'autres élevages - si les élevages ont été tirés au hasard parmi tous les élevages susceptibles d'utiliser ces produits,le facteur élevage est un facteur **aléatoire** et les résultats de cette analyse peuvent être extrapolés aux autres élevages.

Comme nous le verrons plus tard, la méthode d'analyse doit tenir compte de la nature des variables explicatives du modèle.

Remarques

Il est bien évident qu'une équation ne peut pas parfaitement décrire le phénomène étudié. En effet, la réalité est toujours plus complexe que le modèle qui la traduit. C'est pour cette raison que nous compléterons notre modèle en "ajoutant" des variables aléatoires (appelées erreurs ou résidus) qui représenteront nos ignorances et qui synthétisent la variabilité des résultats observés, les imperfections des mesures.Il existe différents type d'erreurs:

1. les erreurs aléatoires
2. les erreurs de mesures
3. les erreurs dues à un mauvais choix de modèle.

Ces trois types d'erreurs sont malheureusement indissociables. On ne peut qu'estimer globalement leur ordre de grandeur. Quelles sont les hypothèses "raisonnables" que l'on peut faire sur ces résidus ?

1. Ils doivent être distribués normalement. En effet, un phénomène suit une loi normale. S'il est la résultante d'un grand nombre d'effets indépendants et petits qui s'ajoutent, cette hypothèse est donc raisonnable.

2. Les erreurs doivent être du même ordre de grandeur quels que soient les facteurs. Ceci veut dire que les résidus pour différents facteurs doivent être à peu près identiques.

3. Ils doivent être indépendants ce qui signifie que l'erreur commise sur une mesure ne doit pas être liée à l'erreur commise sur une autre mesure.

Nous reviendrons plus en détail sur ces hypothèses dans les chapitres consacrés à la régression, et à l'analyse de variance.

Erreur aléatoire - Erreur systématique

Les causes d'erreurs sont nombreuses. Par exemple en Zootechnie, les variations de poids d'un bovin sont considérables au cours d'une journée selon que l'animal vient de boire ou de recevoir sa ration alimentaire. Toutes ces causes de variations entraînent une imprécision, une erreur que l'on qualifiera d'aléatoire. Il ne faut pas que cet ensemble d'erreurs devienne systématique, car à ce moment là, la statistique est impuissante. Aucune méthode ne pourra jamais dire si les différences observées entre trois médicaments sont dues au médicament ou au vétérinaire si chaque médicament a été donné par un vétérinaire différent. Voilà l'exemple de ce qu'il ne faut pas faire ! Pour éviter l'erreur systématique, deux précautions :

- **Les mesures en aveugle**
  L'expérimentateur peut être amené à perdre son objectivité lorsqu'il réalise des mesures traitement par traitement. La solution consiste à travailler en "aveugle" c'est-à-dire sans connaître le traitement affecté à l'unité expérimentale mesurée.

- **La randomisation**
  Randomiser signifie tirer au hasard. Pour constituer des groupes comparables, l'expérimentateur ne doit pas chercher à les équilibrer lui-même, par tâtonnements successifs, mais randomiser. Ainsi, la multitude de facteurs extérieurs qui agissent sur l'erreur expérimentale est répartie d'une manière sensiblement uniforme sur l'ensemble des unités expérimentales.

Classification des analyses

Les analyses des modèles linéaires portent des noms différents selon de la nature des variables explicatives utilisées dans le modèle. Le tableau suivant contient le nom des différentes analyses par nature des variables explicatives

variable(s) explicative(s)	nom de l'analyse
1 quantitative	régression simple
plusieurs quantitatives	régression multiple
plusieurs qualitatives	analyse de variance
1 quantitative et plusieurs qualitatives	analyse de la covariance
plusieurs quantitatives et plusieurs qualitatives	modèle linéaire

En résumé

Un modèle linéaire est une expression qui relie une variable quantitative à des variables (quantitatives et qualitatives).
Les paramètres inconnus interviennent linéairement dans le modèle.
Il est postulé dans le modèle linéaire standard que :
- la variance des observations est constante,
- les observations sont indépendantes,
- les observations sont normalement distribuées.

Chapitre 2

REGRESSION LINEAIRE SIMPLE

C'est le cas le plus simple; l'une des variables (généralement désignée par la lettre x) ne prend que des valeurs choisies a priori (ou du moins déterministes) et l'autre variable (habituellement notée Y), est aléatoire. Le modèle s'écrit sous la forme suivante:

$$Y = \theta_1 + \theta_2 x + \varepsilon$$

où θ_1 et θ_2 sont les paramètres inconnus et ε est le terme d'erreur.

Ou pour être plus précis,

$$Y_i = \theta_1 + \theta_2 x_i + \varepsilon_i,\ i = 1..n \tag{2.1}$$

Si nous reprenons l'exemple sur la taille et l'age des enfants, n représente ici le nombre total d'enfants mesurés, Y_i représente la taille du i^{ieme} enfant x_i l'age du i^{ieme} enfant.

Avant d'aller plus loin dans l'analyse de ce modèle, faisons les hypothèses fondamentales (qu'il ne faudra pas oublier de vérifier).

2.1 Hypothèses

1) les variables aléatoires ε_i sont normalement distribuées
2) les $\left(\varepsilon_i\right)_{i=1..n}$ ont une variance (notée σ^2) identique (on dit souvent que la variance de ε doit être constante)

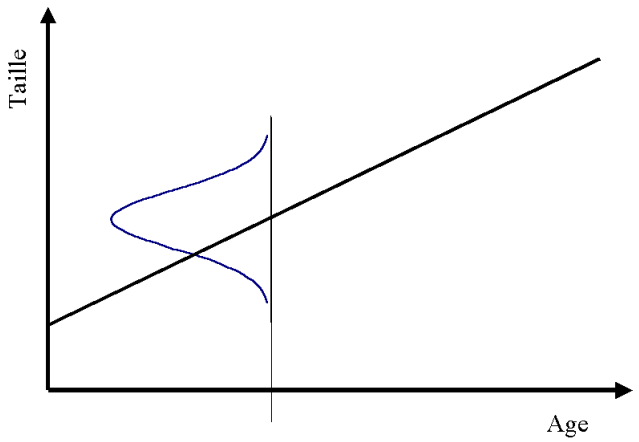

Figure 2.1: Pour un âge donné les tailles sont distribuées normalement

3) leur espérance est nulle,
4) les $\big(\varepsilon_i\big)_{i=1..n}$ sont indépendants.
Les variables x_i étant déterministes, les hypothèses faites sur les $\big(\varepsilon_i\big)_{i=1..n}$ se transposent immédiatement aux $\big(Y_i\big)_{i=1..n}$ de la façon suivante
1') les variables aléatoires $\big(Y_i\big)_{i=1..n}$ sont normalement distribuées
2') les $\big(Y_i\big)_{i=1..n}$ ont une variance (notée σ^2) identique
3') les $\mathbb{E}\big(Y_i\big) = \theta_1 + \theta_2 x_i$
4') les $\big(Y_i\big)_{i=1..n}$ sont indépendantes.
Reprenons l'exemple sur la taille et l'age pour illustrer ces hypothèses:

1) dire que les $\big(Y_i\big)_{i=1..n}$ sont normalement distribuées, signifie que pour un âge (x) fixé si on faisait l'histogramme des tailles sur toute la population des enfants, on trouverait une distribution normale (cf graphique 2.1)

2) si les $\big(Y_i\big)_{i=1..n}$ ont une variance constante alors, la dispersion de la taille à un âge donné, est égale à la dispersion de la taille à un autre âge. Il semble raisonnable de faire cette hypothèse pour les données représentées sur le graphique 2.1. En revanche,le graphique (2.2) illustre une dépendance entre la dispersion des tailles et l'age.

3) Pour chaque x fixé, la moyenne de population des tailles est supposée égale à $\theta_1 + \theta_2 x$. Les points observés sont donc supposés "groupés" (au moins à ε prés) autour d'une droite d'équation $\theta_1 + \theta_2 x$. En d'autres termes, la moyenne de population des écarts entre la taille des individus d'âge x et la

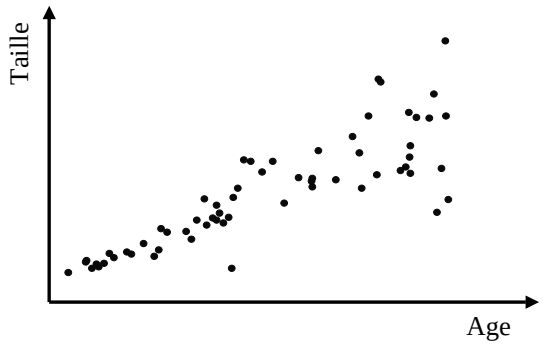

Figure 2.2: la variance des tailles est proportionnelle à l'âge

moyenne de population des tailles des individus du même âge est supposée nulle.

4) la notion de dépendance n'est pas très facile à expliquer sans faire appel aux probabilités ; on peut cependant s'en sortir de façon intuitive. Si la taille est mesurée sur des individus différents, il n'y a aucune raison pour que la taille d'un individu "influence" celle d'un autre individu et dans ce cas l'indépendance peut être postulée. En revanche, supposons que les tailles observées soit celles de plusieurs individus qui auraient été mesurés au cours de leur croissance, la taille d'un individu à 6 ans dépend vraisemblablement de sa taille à 5 ans et dans ce cas, il n'y a aucune raison de postuler *a priori* l'indépendance des Y_i. L'indépendance est une situation confortable pour celui qui analyse les données: les résultats classiques sur les tests ne sont valides que sous cette hypothèse.

2.2 Estimation des paramètres

Le modèle (2.1) contient les deux paramètres inconnus θ_1 et θ_2 et un autre paramètre que nous avons évoqué au paragraphe précédent σ^2.
Ces paramètres sont les paramètres de population (auxquels nous n'avons pas accès), l'objet de ce paragraphe est d'en trouver, à partir de l'échantillon, une approximation raisonnable.

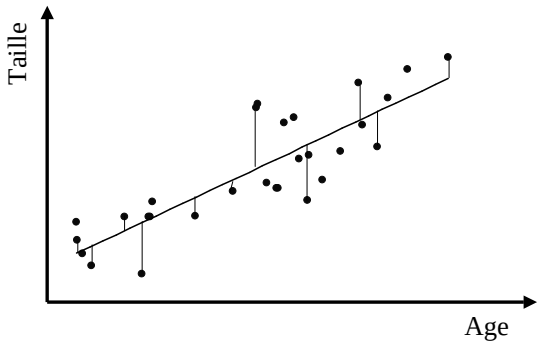

Figure 2.3: la distance est représentée par des barres verticales

A cet effet, deux grandes méthodes sont souvent utilisées:
-**les moindres carrés**
-**le maximum de vraisemblance.**
La méthode des moindres carrés repose sur une idée simple:
on cherche θ_1 et θ_2 de telle façon que la droite $Y = \theta_1 + \theta_2 x$ soit la plus "proche" (dans un sens à préciser) de tous les points observés.
La notion de proximité précédemment évoquée se rapporte à une façon de mesurer la distance entre un point observé (x_i, Y_i) et le point d'abscisse x_i qui se trouve sur la droite (il a donc pour ordonnée $\theta_1 + \theta_2 x_i$. La distance utilisée dans la méthode des moindres carrés est la distance usuelle, celle que nous utilisons tous les jours dans notre espace à trois dimensions (cf graphique 2.3) soit:

$$d^2(Y_i, \theta_1 + \theta_2 x_i) = (Y_i - (\theta_1 + \theta_2 x_i))^2 \tag{2.2}$$

La formule précédente permet de calculer le carré de la distance qui sépare 2 points. Or, nous disposons de n points (dans notre exemple nous disposons de 30 tailles d'enfants donc $n = 30$) : il faut donc construire un indice qui mesurera la distance entre les 30 points mesurés et leur points correspondants sur la droite. La distance utilisée, ou plutôt son carré, est en général notée

SCE (Somme des Carrés des Ecarts) et vaut:

$$SCE = \big(Y_1 - (\theta_1 + \theta_2 x_1)\big)^2 + \big(Y_2 - (\theta_1 + \theta_2 x_2)\big)^2 + ... + \big(Y_n - (\theta_1 + \theta_2 x_n)\big)^2$$
$$= \sum_{i=1}^{n} \big(Y_i - (\theta_1 + \theta_2 x_i)\big)^2$$

On cherche donc la valeur des paramètres θ_1 et θ_2 qui rendra la SCE minimum (on comprend ici la provenance du nom de la méthode).
On peut voir [1] que

$$\hat{\theta_2} = \frac{\sum_{i=1..n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1..n}(x_i - \bar{x})^2} \tag{2.3}$$

et que,

$$\hat{\theta_1} = \bar{y} - \hat{\theta}_2 \bar{x}, \tag{2.4}$$

où $\bar{x}$ et $\bar{y}$ sont respectivement la moyenne des x_i et la moyenne des y_i.
Fidèle à nos conventions du début, le symbole $\hat{\theta}_1$ signifie que θ_1 n'est pas la valeur du paramètre de population, mais une estimation de cette valeur.
La méthode du maximum de vraisemblance repose sur l'idée de FISHER selon laquelle si les données de l'échantillon ont été observées, cela provient du fait que ces données sont les plus vraisemblables. Les estimateurs des paramètres inconnus du modèle sont donc calculés en maximisant une quantité (la vraisemblance) qui "mesure la probabilité d'observer l'échantillon".
Une propriété importante: sous les hypothèses 1), 2), 3) les estimateurs de θ_1 et θ_2 sont exactement les mêmes que ceux que nous avons calculé avec la méthode des moindres carrés (2.3) et (2.4);

un estimateur [2] sans biais de σ est donné par:

$$\hat{\sigma}^2 = \frac{SCE}{n-2} = \frac{1}{n-2} \sum_{i=1}^{n} \big(Y_i - (\theta_1 + \theta_2 x_i)\big)^2.$$

Nous disposons maintenant d'estimations des quantités θ_1 et θ_2, et il est bien évident que si un autre échantillon avait été considéré, nous aurions

[1] Il suffit de dériver la SCE par rapport à θ_1 et θ_2

[2] Les estimateurs calculés avec la méthode du maximum de vraisemblance sous les hypothèses 1), ... 4 sont des estimateurs sans biais, de variance minimum, efficaces, en d'autres termes, la méthode du maximum de vraisemblance "est ce qui se fait de mieux".

obtenu d'autres valeurs de $\hat{\theta}_1$ et de $\hat{\theta}_2$.
Ceci illustre le fait que les estimateurs $\hat{\theta}_1$ et $\hat{\theta}_2$ sont aléatoires. On peut montrer que sous les hypothèses 1), 2), 3), 4) ces estimateurs suivent une loi normale. En particulier:

$$\hat{\theta_1} \sim N(\theta_1, \sigma^2(\frac{1}{n} + \frac{\bar{x}^2}{\sum(x_i-\bar{x})^2})) \quad (2.5)$$
$$\hat{\theta_2} \sim N(\theta_2, \frac{\sigma^2}{\sum(x_i-\bar{x})^2}) \quad (2.6)$$

Les variances des estimateurs dépendent de la variance de population σ^2 estimée par $\hat{\sigma^2}$. Nous appellerons estimateur de la variance de l'estimateur la quantité obtenue en remplaçant dans (2.5) et (2.6) σ^2 par $\hat{\sigma^2}$. Les anglo-saxons nomment en général **standard error (s.e.)** la racine carrée de la variance estimée des estimateurs. Ces quantités sont utilisées pour construire des intervalles de confiance et faire des tests. Pour chaque valeur de x (age) nous disposons d'une valeur de Y (taille) prédite (on dit aussi ajustée) que nous noterons $\hat{Y} = \hat{\theta}_1 + \hat{\theta}_2 x$. On appelle **résidu** (ils seront noté $\hat{\varepsilon}$) la différence entre les valeurs de Y observées et les valeurs ajustées soit:

$$\hat{\varepsilon} = Y - \hat{Y} = Y - (\hat{\theta}_1 + \hat{\theta}_2 x)$$

ou encore:

$$\hat{\varepsilon}_i = Y_i - \hat{Y}_i = Y_i - (\hat{\theta}_1 + \hat{\theta}_2 x_i) \text{ pour } i = 1..n$$

Le paragraphe suivant est consacré aux tests sur le paramètres et à la construction d'intervalles de confiance.

2.3 Test sur les paramètres

Une grande variété de tests peut être réalisée en régression, il dépendent de la question à laquelle on veut répondre. Notre objectif n'étant pas d'être exhaustif, nous nous contenterons de rappeler les tests usuels sur les paramètres. Rappelons le modèle avec lequel nous allons travailler:

$$Y = \theta_1 + \theta_2 x + \varepsilon$$

Le paramètre θ_1 représente ici l'ordonnée à l'origine de la droite $y = \theta_1 + \theta_2 x$ (pour s'en convaincre, il suffit de prendre $x = 0$) alors que le paramètre θ_2 représente la pente de la droite.
Si $\theta_1 = 0$ alors la droite passe par 0, si $\theta_2 = 0$, alors pour tout x, y est constant, ou encore en utilisant notre exemple, $(\theta_2 = 0) \Longrightarrow$ quel que soit l'âge, la taille est identique.
test sur θ_1 Pour tester :$H_0 : \theta_1 = \theta_{1_0}$ contre $H_1 : \theta_1 \neq \theta_{1_0}$, on utilise la statistique de test:

$$T = \frac{|\hat{\theta}_1 - \theta_{1_0}|}{s.e._{\theta_1}}$$

avec $s.e._{\theta_1} = \sqrt{\hat{\sigma}^2(\frac{1}{n} + \frac{\bar{x}^2}{\sum(x_i - \bar{x})^2})}$ et la règle de décision est :
si $T > t^{n-2}_{1-\alpha/2}$ alors on rejette H_0 avec un risque α
$t^{n-2}_{1-\alpha/2}$ est la valeur limite au seuil $1 - \frac{\alpha}{2}$ d'une loi de *Student* à $n-2$ degrés de liberté. [3]
test sur θ_2
Pour tester :$H_0 : \theta_2 = \theta_{2_0}$ contre $H_1 : \theta_2 \neq \theta_{2_0}$, on utilise la statistique de test:

$$T = \frac{|\hat{\theta}_2 - \theta_{2_0}|}{s.e._{\theta_2}}$$

avec $s.e._{\theta_2} = \sqrt{\frac{\hat{\sigma}^2}{\sum(x_i - \bar{x})^2}}$ et la règle de décision est :
si $T > t^{n-2}_{1-\alpha/2}$ alors on rejette H_0 avec un risque α
$t^{n-2}_{1-\alpha/2}$ est la valeur limite au seuil $1 - \frac{\alpha}{2}$ d'une loi de *Student* à $n-2$ degré de liberté. [4] Nous verrons un exemple numérique complet à la fin de ce chapitre.

2.4 Intervalles de confiance

Intervalle de confiance des paramètres

Le paragraphe (2.2) était consacré à l'estimation des paramètres du modèle. Ainsi, chaque paramètre est estimé par *une* valeur; on parle dans ce cas

[3] Les logiciels de statistique testent systématiquement les hypothèses $H_0 : \theta_1 = 0$ contre $H_1 : \theta_1 \neq 0$

[4] comme pour θ_1, les logiciels de statistique testent les hypothèses $H_0 : \theta_2 = 0$ contre $H_1 : \theta_2 \neq 0$

d'**estimation ponctuelle**. Souvent, l'information désirée est d'un autre ordre, il s'agit de connaître *les* valeurs que peut raisonnablement prendre ce paramètre. On parle alors d'**estimation par intervalle** ou encore d'**intervalle de confiance**.

Les intervalles de confiance de sécurité $1-\alpha$ des paramètres θ_1 et θ_2 sont donnés par :

(2.7)

$$\hat{\theta}_1 - t^{n-2}_{1-\alpha/2}\sqrt{\hat{\sigma}^2(\frac{1}{n}+\frac{\bar{x}^2}{\sum(x_i-\bar{x})^2})} \leq \theta_1 \leq \hat{\theta}_1 + t^{n-2}_{1-\alpha/2}\sqrt{\hat{\sigma}^2(\frac{1}{n}+\frac{\bar{x}^2}{\sum(x_i-\bar{x})^2})}$$

(2.8)

$$\hat{\theta}_2 - t^{n-2}_{1-\alpha/2}\sqrt{\frac{\hat{\sigma}^2}{\sum(x_i-\bar{x})^2}} \leq \theta_2 \leq \hat{\theta}_2 + t^{n-2}_{1-\alpha/2}\sqrt{\frac{\hat{\sigma}^2}{\sum(x_i-\bar{x})^2}}$$

Remarque : Ces formules apparaissent comme très compliquées, notons cependant qu'elles peuvent être réécrites sous la forme suivante:

$$\hat{\theta}_1 - t^{n-2}_{1-\alpha/2} s.e._{\theta_1} \leq \theta_1 \leq \hat{\theta}_1 + t^{n-2}_{1-\alpha/2} s.e._{\theta_1}$$

et pour θ_2 on obtient:

$$\hat{\theta}_2 - t^{n-2}_{1-\alpha/2} s.e._{\theta_2} \leq \theta_2 \leq \hat{\theta}_2 + t^{n-2}_{1-\alpha/2} s.e._{\theta_2}$$

La *structure* de la construction de l'intervalle apparaît ainsi plus clairement.

Ces intervalles de confiance sont symétriques autour de $\hat{\theta}_1$ et $\hat{\theta}_2$. Il est possible de construire des intervalles non symétriques, mais ils seront plus larges. Les intervalles de confiance symétriques sont les intervalles de largeur minimum.

Si α augmente, la largeur de l'intervalle diminue. Par exemple un intervalle de confiance à 95% ($\alpha = 5\%$) est plus large qu'un intervalle de confiance à 90% ($\alpha = 10\%$).

Si n (effectif de l'échantillon) augmente, la largeur de l'intervalle de confiance diminue: Plus on a d'informations plus il est possible d'être précis sur les paramètres.

En comparant la formule de l'intervalle de confiance des paramètres avec la formule des tests sur ces paramètres on s'aperçoit que tester les hypothèses $H_0 : \theta_2 = \theta_{2_0}$ contre $H_1 : \theta_2 \neq \theta_{2_0}$ avec un risque de première espèce de α

est strictement équivalent à regarder si θ_{2_0} est en dehors de l'intervalle de confiance de θ_2 de sécurité $1-\alpha$. Ceci est aussi vrai pour θ_1.
Passons maintenant à d'autres intervalles de confiances très utiles pour la prévision.

Intervalle de confiance de la droite de régression (de la moyenne)

La droite de régression dont nous disposons ($\hat{Y} = \hat{\theta}_1 + \hat{\theta}_2 x$) est construite avec des quantités aléatoires ($\hat{\theta}_1$ et $\hat{\theta}_2$) dont la valeur varie avec l'échantillon. La droite que nous avons observée avec cet échantillon n'est donc qu'un représentant de toute les droites que nous aurions pu observer en prenant différents échantillons. L'objet de l'intervalle de confiance à $1-\alpha\%$ de la droite de régression est de donner des limites entre lesquelles on va trouver $(1-\alpha)\%$ des droites possibles. En voici la formule:

$$\hat{\theta}_1+\hat{\theta}_2 x - t^{n-2}_{1-\alpha/2}\sqrt{\hat{\sigma}^2(\frac{1}{n}+\frac{(x-\bar{x})^2}{\sum(x_i-\bar{x})^2})} \leq \theta_1+\theta_2 x \leq \hat{\theta}_1+\hat{\theta}_2 x + t^{n-2}_{1-\alpha/2}\sqrt{\hat{\sigma}^2(\frac{1}{n}+\frac{(x-\bar{x})^2}{\sum(x_i-\bar{x})^2})}$$

ou encore

$$\hat{Y}(x) - t^{n-2}_{1-\alpha/2}\sqrt{\hat{\sigma}^2(\frac{1}{n}+\frac{(x-\bar{x})^2}{\sum(x_i-\bar{x})^2})} \leq Y(x) \leq \hat{Y}(x) + t^{n-2}_{1-\alpha/2}\sqrt{\hat{\sigma}^2(\frac{1}{n}+\frac{(x-\bar{x})^2}{\sum(x_i-\bar{x})^2})}.$$

En faisant varier x on obtient ainsi deux courbes.
On peut remarquer qu'en prenant $x = 0$, on retrouve bien l'intervalle de confiance de l'ordonnée à l'origine θ_1.

Intervalle de confiance d'une valeur prédite : intervalle de prédiction

La régression peut servir à prédire (pas au sens magique du terme). Par exemple, on veut connaître pour un âge donné (x_0) les tailles vraisemblables. La droite de régression nous donne déjà une information sur la taille moyenne à l'age x_0. Mais comme nous venons de le voir, la droite est soumise à des variations aléatoires (on ne dispose que d'un échantillon), la taille prédite avec cette droite sera donc soumise à ces mêmes variations aléatoires. Il reste une dernière source de variation de cette valeur prédite qui, elle, n'est pas due

à l'estimation mais qui est due au fait que, pour un âge donné, plusieurs tailles (dans la population) sont possibles. Ceci explique la ressemblance des formules de l'intervalle de confiance de la droite de régression et de celui d'une valeur prédite. L'intervalle qui contient la taille de $100(1-\alpha)\%$ des individus d'âge x est donnée par

$$\begin{aligned}\hat{\theta}_1 + \hat{\theta}_2 x_0 - \qquad & t^{n-2}_{1-\alpha/2}\sqrt{\hat{\sigma}^2(1+\tfrac{1}{n}+\tfrac{(x_0-\bar{x})^2}{\sum(x_i-\bar{x})^2})} \leq Y(x_0) \\ & \leq \hat{\theta}_1 + \hat{\theta}_2 x_0 + t^{n-2}_{1-\alpha/2}\sqrt{\hat{\sigma}^2(1+\tfrac{1}{n}+\tfrac{(x_0-\bar{x})^2}{\sum(x_i-\bar{x})^2})}.\end{aligned}$$

Remarque 2.4.1 *On peut remarquer que pour $x_0 = \bar{x}$ les deux intervalles précédents ont une largeur minimale. C'est en ce point que la taille moyenne est estimée avec la plus grande précision.*
Le terme $\sum(x_i - \bar{x})^2$ qui intervient dans les intervalles de confiance est proportionnel à la variance empirique des x_i. Plus les x_i sont dispersés, plus ce terme est grand et plus les intervalles de confiance sont étroits. Une façon optimale de planifier l'expérience (choisir les x_i avant l'expérience) consiste donc à prendre des x_i les plus éloignés possibles. Ce plan d'expérience est optimal dés lors qu'il est avéré que la droite est effectivement la courbe qui décrit le mieux les variations moyennes des Y en fonction de x. Dans le cas contraire, il est souhaitable de répartir les x_i afin d'être en mesure de contrôler les écarts au modèle.

2.5 Vérification des hypothèses

Comme nous l'avons vu dans les paragraphes précédents tous les résultats (estimation,tests, intervalles de confiance) de ce chapitre reposent sur les hypothèses fondamentales. Les résidus de la régression sont des "outils" privilégiés pour vérifier ces hypothèses.
Voici un plan de vérification qu'il serait bon de suivre après chaque régression (les hypothèses à vérifier sont classées par ordre d'importance décroissante).
1) Vérification de l'**homoscédasticité** (la variance doit être constante),
1') Vérification de l'**homoscédasticité** (la variance doit être constante),
2) Vérification de **l'indépendance** des observations,
3) Vérification de la **normalité** des observations.

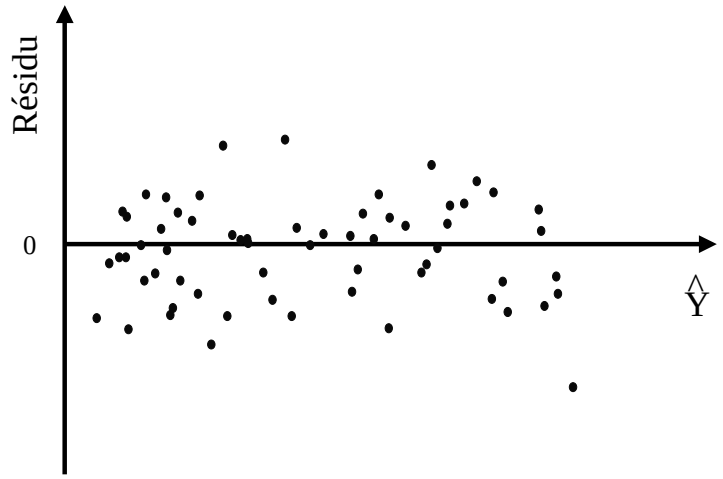

Figure 2.4: La dispersion des résidus autour de zéro semble constante, l'hypothèse d'homoscédasticité est raisonnable.

Vérification de l'homoscédasticité

Cette hypothèse est la plus importante de toute. Une forte violation de cette dernière entraîne des conséquences désastreuses sur :
- les standard error ($s.e.$) des paramètres
- les risques des tests
- les intervalles de confiance.

La méthode la plus couramment utilisée est la vérification graphique: elle consiste à représenter les résidus en fonction des valeurs ajustées, des valeurs observées ou des valeurs de x. Il n'est pas possible d'étudier en détails tous les cas possibles. Les graphiques ci-dessous illustrent les cas les plus courants:

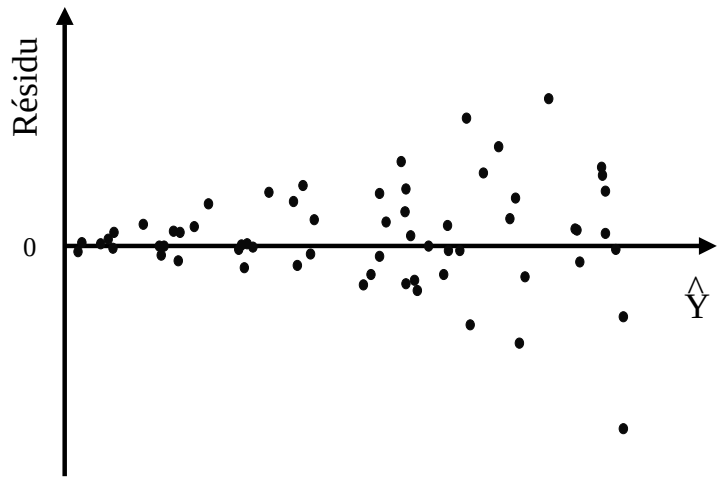

Figure 2.5: La variance des résidus augmente avec $\hat{Y}$.

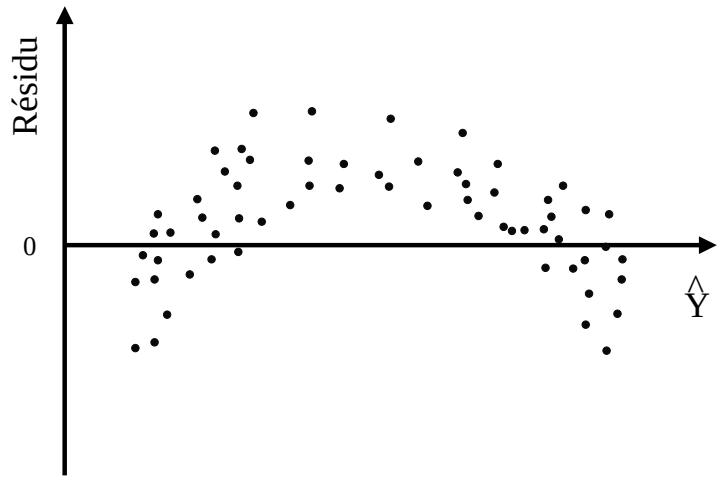

Figure 2.6: Dans cet exemple, l'hypothèse 1' n'est pas raisonnable. Il reste de la "structure", le modèle surestime les faibles et fortes valeurs de Y et sous estime les valeurs moyennes : le modèle de régression linéaire simple n'est pas suffisant, il faut rajouter des variables.

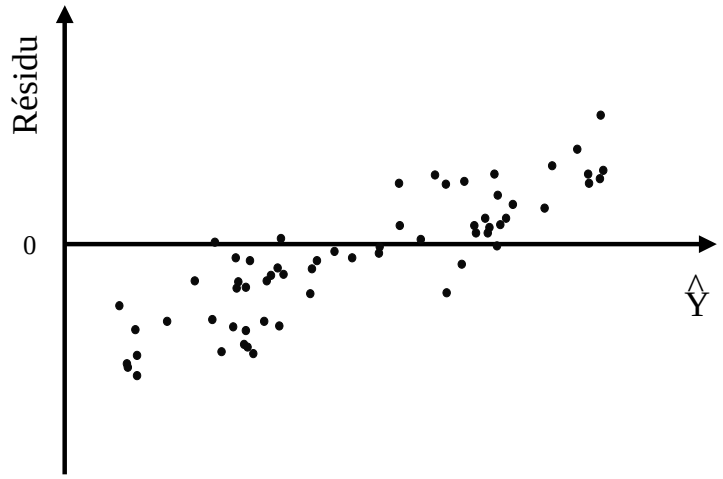

Figure 2.7: Ici encore l'hypothèse 1' n'est pas raisonnable. Le modèle est insuffisant car il surestime les faibles valeurs de Y et sous estime les fortes valeurs de Y. Il reste donc de la "structure".

Vérification de l'indépendance

Quand la régression est réalisée sur des données qui varient au cours du temps, les observations peuvent ne pas être indépendantes. Pour vérifier l'indépendance, un test est habituellement utilisé: le test de **Durbin Watson**.

Il est basé sur la statistique:

$$d = \frac{\sum_{i=2}^{n}(e_i - e_{i-1})^2}{\sum_{i=1}^{n} e_i^2}$$

Les $(e_i)_i$ sont les résidus de la régression et n est le nombre d'observations.

On peut montrer que $0 \leq d \leq 4$ et que

$$\begin{aligned} d &\approx 2 - 2\frac{\sum_{i=2}^{n}(e_i e_{i-1})^2}{\sum_{i=1}^{n} e_i^2} \\ &\approx 2 - 2\rho_e \end{aligned}$$

ρ_e est le coefficient d'autocorrelation d'ordre 1 des résidus.

Il est obtenu en calculant la corrélation entre la série des résidus et la même

série décalée de 1.
Si $\rho_e = 0 \Longrightarrow d \approx 2$ alors les résidus sont non corrélés
Si $\rho_e \neq 0 \Longrightarrow d \neq 2$ les résidus sont corrélés.

Vérification de la normalité

Cette étape n'est pas aussi importante qu'on le croit généralement. La normalité est une propriété qui permet aux estimateurs de converger rapidement. Le *théorème central limit* [5] nous assure que pour des échantillons assez grands, les estimateurs que nous utilisons sont normalement distribués. La symétrie des distributions observées est un critère important qui assure une convergence rapide vers la loi normale.
Certaines méthodes utilisées pour vérifier la normalité ont déjà été vues ; elles ne seront donc pas détaillées dans ce paragraphe. On peut utiliser, entre autres, des méthodes graphiques (PPLOT, histogramme, boxplot, stem and leaf) et des tests (χ^2, Kolmogorov-Smirnov...)

2.6 Détection et élimination des valeurs aberrantes

Un examen critique des données est une étape importante en statistique. Il existe deux grands types de données généralement classée comme aberrantes :
- les données qui ne sont pas habituelles
- les données qui violent une hypothèse de l'analyse statistique utilisée
Différentes attitudes devraient être adoptées suivant la nature du problème rencontré.
Les données provenant d'erreurs grossières de mesures ou d'erreurs de frappe doivent être supprimées de l'analyse.
Seul un jugement biologique permet de déclarer une valeur comme aberrante. Souvent, après un examen attentif des données on trouve des valeurs inhabituelles.
Un expérimentateur prudent doit alors rechercher la (les) cause(s) de telles

[5] La version du TCL que nous avons vu au cours du premier module ne "fonctionne" que sous l'hypothèse d'indépendance, il existe d'autres versions qui montrent la convergence en loi quand cette hypothèse n'est pas trop violée

valeurs. Deux cas de figures se présentent alors:
- soit la cause est identifiée et il faut changer la donnée ou la méthode d'analyse.
- soit la cause n'est pas identifiée et un test statistique peut être utilisé pour détecter une valeur aberrante (la prudence doit être la règle principale).
L'examen graphique des résidus est une méthode couramment employée pour identifier les données suspectes (graphique $e * \hat{Y}$). Une autre technique consiste à calculer des indices pour chaque résidu. La plupart des indices calculés par les logiciels de statistique ont une signification inférentielle. Les trois les plus couramment usités sont:
-les résidus studentisés
-les résidus de cook
-les contributions
Les **résidus studentisés** s'obtiennent en calculant pour chaque résidu son standard error (s.e.) et en divisant chaque résidu par cette quantité.
Pour être plus précis, notons $s.e_i$ le standard error du i^{ieme} résidu, le résidu studentisé est alors défini par $st_i = \frac{e_i}{s.e_i}$. Sous l'hypothèse que $e_i \sim N(0, \sigma^2)$, la quantité st_i suit une loi de *Student* à $n - p - 2$ degrés de liberté.
p représente le nombre de paramètres indépendants estimés dans le modèle (y compris la constante) . Il est donc possible de tester chaque l'"aberrance" de chaque résidu en utilisant un test de student.
D'autres indices sont aussi utilisés:les **contributions** (leverage) et **cook** mesurent la contribution de chaque résidu à la variance résiduelle (non expliquée par le modèle).
Sous les hypothèses usuelles (hypothèses 1) 2) 3)) les résidus de **cook** [6] suivent une loi de Fisher à p et $n - p$ degrés de liberté. Une méthode pour identifier les observations qui contribuent trop à la variance résiduelle consiste donc à réaliser un test de Fisher sur le résidu de cook (i.e. de comparer sa valeur à la valeur limite à un seuil donné d'une loi de Fisher à p et n-p ddl).
Pour des données Gaussiennes, les leverage devraient être voisines de $\frac{p}{n}$; p représente le nombre de paramètres indépendants estimés dans le modèle (y compris la constante). Si pour un résidu,le leverage correspondant est

[6] Les résidus de cook ne sont calculés que pour la régression linéaire simple, dans le cas de régression multiple, on travaillera sur le leverage

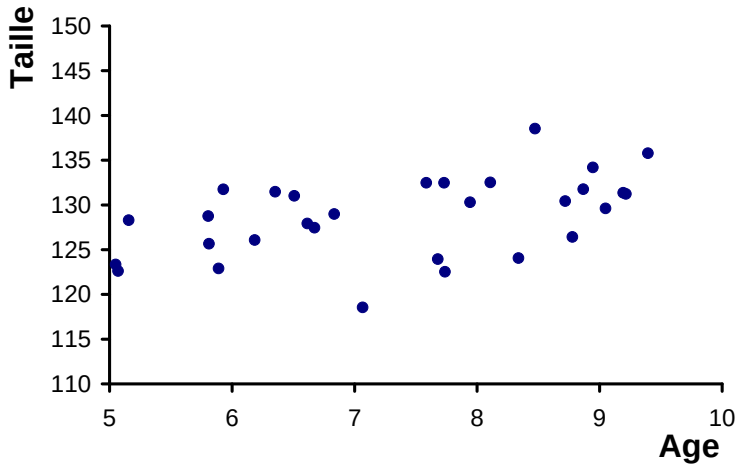

Figure 2.8: La taille apparaît liée linéairement à l'age.

supérieur à $\frac{2p}{n}$, la donnée peut être considérée comme suspecte.

2.7 Un exemple

Estimation et tests

Etude de la relation entre taille et âge sur des enfants entre 5 et 10 ans. Nous disposons de 30 données. Le graphique 2.8 représente les résultats obtenus.

Nous allons donc utiliser le modèle suivant:

$$TAILLE = \theta_1 + \beta age + \varepsilon$$

La regression fournit les résultats suivants :

Regression

Dep var: TAILLE N: 30 Multiple R: .473 Squared Multiple R: .223
Adjusted Squared Multiple R: .196 Standard Error of Estimate: 4.374

Variable	Coefficient	Std Error	Tolerance	T	P(2 tail)
CONSTANT	118.352	3.653	.	32.399	0.000
age	1.406	0.495	.100E+01	2.838	0.008

Analysis of Variance

Source	Sum-of-Squares	DF	Mean-Square	F-Ratio	P
Regression	154.113	1	154.113	8.055	0.008
Residual	535.709	28	19.132		

Durbin-Watson D Statistic 1.968
First Order Autocorrelation .004

Détaillons un peu la sortie : La variable dépendante (*Dep var*) est la *TAILLE* Nous disposons de *N=30* données.

Le **coefficient de corrélation** (R) entre *TAILLE* et *âge* vaut 0.473 (c'est en général assez difficile à interpréter) en revanche, son carré (R^2=*Squared Multiple R*) appelé coefficient de détermination s'interprète comme la part de dispersion expliquée par le modèle ici $0.223(= 22.3\%) = \frac{154.113}{535.709 + 154.113}$.

Le coefficient de détermination est calculé à partir de l'échantillon que nous avons observé. Sa valeur sur estime ce que nous obtiendrions si nous utilisions un autre échantillon. Le coefficient de détermination ajusté *Adjusted Squared Multiple R* est la valeur qu'il faut s'attendre à trouver si nous utilisons le modèle avec un autre échantillon de taille n. Sa valeur est ici .223.

Il se calcule à partir du coefficient de détermination en utilisant la formule suivante:

$$\begin{aligned}\text{Adjusted Squared Multiple R:.196} \quad &= 1 - (1 - R^2)\tfrac{(n-1)}{n-p} \\ &= 1 - (1 - (0.223))\tfrac{(30-1)}{30-2}\end{aligned}$$

La variance résiduelle (non expliquée par le modèle) vaut :19.132) l'écart-type résiduel (*Standard Error of Estimate*) vaut donc:$4.374 = \sqrt{19.132}$ Les lignes suivantes sont consacrées à l'étude des paramètres inconnus du modèle

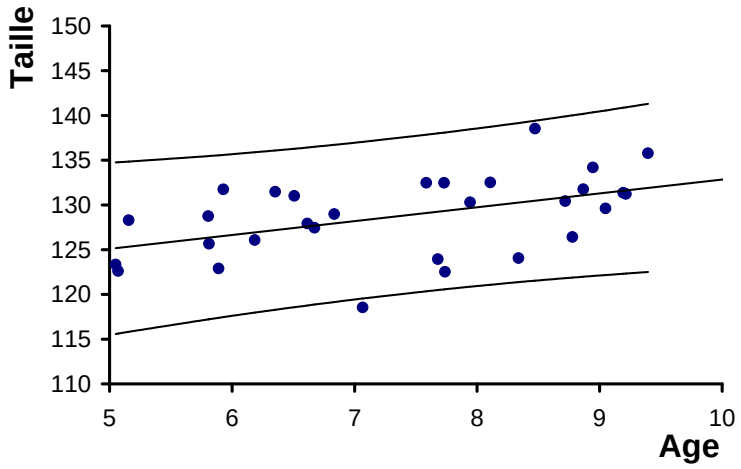

Figure 2.9: La droite $\hat{TAILLE} = 118.352 + 1.406 * AGE$ et son intervalle de confiance à 95%.

ainsi: $\hat{\theta_1} = 118.352$ et $\hat{\beta} = 1.406$, nous en déduisons l'équation de la droite de régression :

$$\hat{TAILLE} = 118.352 + 1.406 * AGE \tag{6.9}$$

Le graphique 2.9 représente la droite de régression d'équation (6.9) ainsi que son intervalle de confiance à 95%.

L'écart type de $\hat{\theta_1}$ et de $\hat{\beta}$ (*Std Error*) que nous avons noté $s.e._{\theta_1}$ et $s.e._{\beta}$ valent respectivement 3.653 et 0.495.
On peut maintenant construire des intervalles de confiance pour ces estimateurs. Par exemple, un intervalle de confiance à 95% de β se calcule de la façon suivante:

$$\hat{\beta} - t^{n-2}_{1-\alpha/2} s.e._{\beta} \leq \beta \leq \hat{\beta} + t^{n-2}_{1-\alpha/2} s.e._{\beta}$$

soit:

$$\begin{aligned} 1.406 - 2.048 * 0.495 \quad &\leq \beta \leq 1.406 + 2.048 * 0.495 \\ 1.406 - 1.00 \qquad &\leq \beta \leq 1.406 + 1.00 \end{aligned}$$

avec $t^{n-2}_{1-\alpha/2} = t^{28}_{1-0.05/2} = 2.048$ (cette valeur provient d'une table de Student)
Le coefficient intitulé *TOLERANCE* en face de chaque paramètre mesure la corrélation entre la variable devant laquelle se trouve le paramètre et les autres variables du modèle. il n'est d'aucun intérêt en régression linéaire simple, nous verrons en régression multiple comment le calculer et l'interpréter.

Les valeurs T, sont les statistiques utilisées pour les tests de *Student* sur les paramètres θ_1 resp(β).Les hypothèses testées sont: $H_0 : \theta_1 = 0$ contre $H_1 : \theta_1 \neq 0$ resp $\big(H_0 : \beta = 0$ contre $H_0 : \beta \neq 0\ \big)$.
On en déduit donc que pour le test
$H_0 : \theta_1 = 0$ contre $H_1 : \theta_1 \neq 0$, $T = \frac{118.352}{3.653} = 32.399$ et pour le test $H_0 : \beta = 0$ contre $H_0 : \beta \neq 0$, $T = \frac{1.406}{0.495} = 2.838$

Ces valeurs devraient être comparés à la valeur limite d'une loi de Student à $n-2$ degrés de liberté; cependant, la plupart des logiciels fournissent un risque $P(2tail)$ qui correspond au risque de première espèce observé pour ces tests d'hypothèses. Dans la mesure où ce risque dépend de l'échantillon, nous ne l'utilisons pas tel quel : nous le comparons au risque α **fixé a priori**. Ainsi, il y a un risque de $< 1\%$ de dire que $\beta \neq 0$ alors qu'en réalité $\beta = 0$..
Il est bien évident que de tel test ne répondent pas toujours à la question que l'on se pose. Supposons par exemple que la question posée soit: *la droite a t- elle une pente différente de 0.5* ou encore:$H_0 : \beta = 0.5$ ou $H_1 : \beta \neq 0.5$.
Pour tester cette hypothèse, il faut construire la statistique adéquate, soit :

$$T = \frac{|1.406 - 0.5|}{0.495} = 1.830$$

à comparer à la valeur $t^{28}_{1-0.05/2} = 2.048$.
On ne rejette donc pas l'hypothèse $\beta = 0.5$ ($1.830 < 2.048$).
Passons maintenant au tableau intitulé **Analysis of Variance**.
On trouve ici l'effet des différentes sources de variations. Rappelons que $Var(Y) = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2$ estime la variance des Y_i et c'est précisément cette dispersion que nous voulons expliquer par la régression.
A cet effet remarquons que $SCE_t = \sum_{i=1}^{n}(Y_i - \bar{Y})^2$ mesure la "quantité d'information" totale contenue dans les données.
Afin d'apprécier l'efficacité de la régression on peut décomposer cette "quantité d'information totale" en une "quantité d'information" expliquée par la

régression et une ł"quantité d'information" non expliquée par le modèle (la régression) soit :

$$\begin{aligned} SCE_t &= \sum_{i=1}^{n}(Y_i - \bar{Y})^2 \\ &= \Big(\sum_{i=1}^{n}(Y_i - \bar{Y})^2 - \sum_{i=1}^{n}(Y_i - \hat{Y})^2\Big) + \sum_{i=1}^{n}(Y_i - \hat{Y})^2 \\ &= SCE_{reg} + SCE_{res} \end{aligned}$$

La quantité $SCE_{reg} = Sum - of - Squares_{Regression} = 154.113$ est la "quantité d'information" expliquée par le modèle,tandis que $SCE_{res} = Sum - of - Squares_{Residual} = 535.709$ est la "quantité d'information " non expliquée par le modèle [7].
Le degré de liberté DF de la régression vaut $p - 1 = 2 - 1$ et le degré de liberté de la SCE résiduelle (déjà notée SCE_{res}) vaut n-p.
Nous reviendrons dans le chapitre consacré à l'analyse de variance sur cette notion de degré de liberté.

Les Carrés Moyens *Mean-Square* s'obtiennent en divisant chaque SCE par le degré de liberté correspondant soit: $154.113 = \frac{154.113}{1}$ et $\frac{19.132=535.709}{28}$.
Ces quantités sont d'authentiques variances: ainsi, 154.113 est la **variance expliquée par la régression** et 19.132 est la **variance résiduelle** (non expliquée par la régression).
F-Ratio est la statistique utilisée pour faire un test de FISHER, elle se calcule en faisant le quotient de la variance expliquée par l'effet à tester (ici la régression) par la variance résiduelle soit: $F = 8.055 = \frac{154.113}{19.132}$.
Comme dans tout test de FISHER, on compare ce F à la valeur limite d'une loi de FISHER avec comme premier degré de liberté, le degré de liberté du numérateur (ici=1) et comme second degré de liberté, le degré de liberté du dénominateur (ici 28).

La quantité P correspond au risque observé de 1^{iere} espèce (ici 0.008). Le test F est donc significatif à 5%, on en déduit que le modèle (dans sa

[7] On peut noter que cette quantité est précisément celle que nous avons minimisée en θ_1 et β

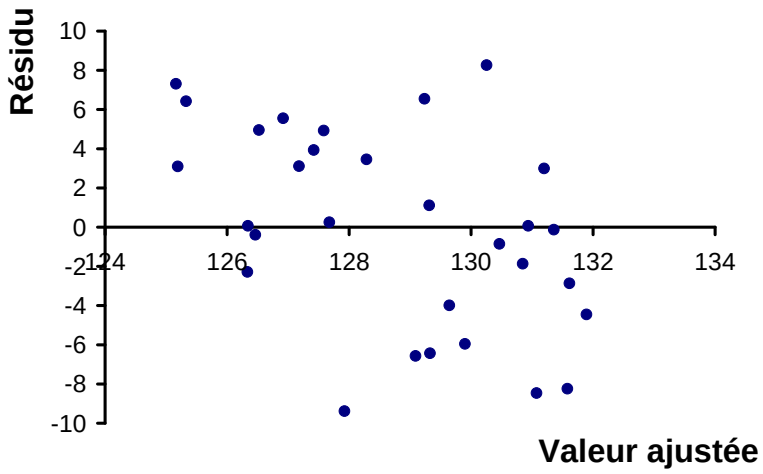

Figure 2.10: Les résidus se répartissent de façon homogène autour de zéro, l'hypothèse d'homoscédasticité ne semble pas violée.

globalité) explique significativement des variations de la variable $TAILLE$.
Remarque : On peut remarquer que les risques du "F" et du test T associé à β sont identiques. Ce n'est pas un hasard. En effet, dans la cas particulier de la régression linéaire simple, le seul facteur de variation pour expliquer la taille est l'age ; il est donc tout a fait normal que le test sur β (qui mesure l'influence de l'age sur la taille) donne le même résultat que le test sur la régression tout entière. En fait, dans ce cas particulier, T et F sont liés par la relation:$T^2 = (2.838)^2 = F = 8.055$.

Examen des résidus

Vérifions tout d'abord l'homoscédasticité. Le graphique 2.10 représente les résidus en fonction des valeurs ajustées. Voyons maintenant comment se comportent les résidus en fonction de la variable AGE. Le graphique 5.1 montre que les résidus se répartissent de façon homogène autour de zéro, l'hypothèse d'homoscédasticité ne semble pas violée. La vérification de non dépendance des observations peut être réalisée avec la statistique de **Durbin**

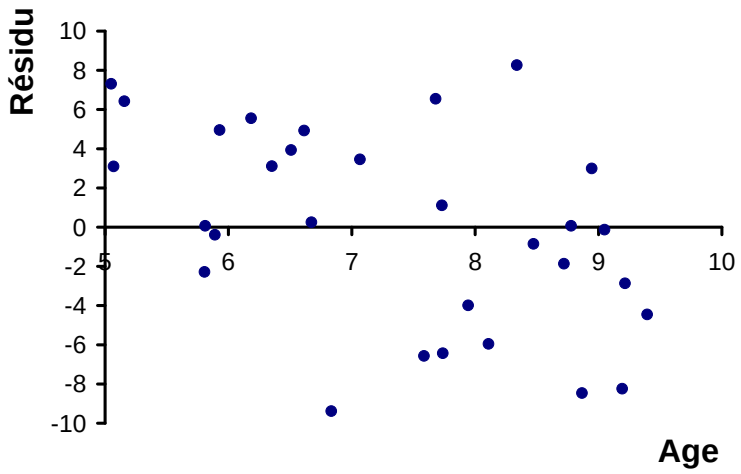

Figure 2.11: Les résidus se répartissent de façon à peu prés homogène autour de 0. L'hypothèse d'homoscédasticité est ici raisonnable.

Watson qui ici vaut 1.968. La corrélation entre les résidus est donc très faible. le graphique 5.2 représente la droite de Henry (PPLOT) des résidus. En résumé, nous n'avons fortement transgressé aucune hypothèse (c'est un cas d'école).

2.8 Non respect des hypothèses

Les hypothèses ne peuvent pas être trop violées sans quelques désagréments sur la validité des résultats de l'analyse statistique sont:
- l'homoscédasticité
- l'indépendance des observations
- la normalité des observations.

Non-linéarité

Il ne faut cependant pas oublier que le point de départ de notre étude de la régression est la linéarité du modèle. Il arrive souvent que cette hypothèse

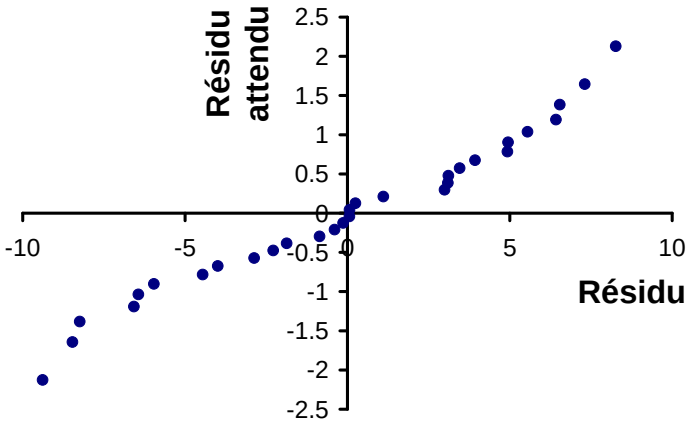

Figure 2.12: Mis à part les effets de bords inévitables dans ce genre de graphique, les points semblent bien alignés. L'hypothèse de normalité semble donc raisonnable.

implicitement faite, ne soit pas vérifiée (un simple examen des résidus vous montre que la structure linéaire n'est pas adaptée à vos données). Les exemples sont légions, on peut penser par exemple aux relations allométriques entre deux mensurations Y et x qui souvent sont de la forme:$Y = x^{\beta}$
Ou encore à des modèle de pharmacocinétique qui souvent peuvent s'exprimer (au moins localement dans le temps) sous la forme: $Y = \alpha e^{\beta x}$ ou $Y = \frac{x}{\alpha x - \beta}$.
Pour se ramener au modèle linéaire classique, on peut, dans certains cas , faire des transformations de variables. Sans vouloir citer de façon exhaustive toutes les relations qui peuvent être linéarisées en voici un certain nombre :

fonction	transformation	forme linéaire	graphique
$Y = \alpha x^{\beta}$	$Y' = \ln Y$	$Y' = \ln \alpha + \beta x'$	1
	$x' = \ln x$		
$Y = \alpha e^{\beta x}$	$Y' = \ln Y$	$Y' = \ln \alpha + \beta x$	2
$Y = \dfrac{x}{\alpha x - \beta}$	$Y' = \frac{1}{Y}$	$Y' = \alpha - \beta x'$	3
	$x' = \frac{1}{x}$		
$Y = \dfrac{e^{\alpha+\beta x}}{1 + e^{\alpha+\beta x}}$	$Y' = \ln \frac{Y}{1-Y}$	$Y' = \alpha + \beta x$	4

Si pour des raisons pratiques (de nature biologique), l'erreur dans le modèle initial (non linéaire) est additive, la transformation détruit complètement la structure de cette erreur; la conséquence souvent fâcheuse est la perte d'homoscédasticité. Il faut noter de plus que les estimateurs obtenus en linéarisant le modèle sont en général biaisés (mais souvent asymptotiquement sans biais).

Par conséquent, quand il ne s'agit que de décrire les données, l'approximation linéaire (la linéarisation) est une bonne technique, en revanche, dés lors qu'il s'agit d'analyser finement les données avec l'espoir de faire de la prévision ou de l'inférence, la linéarisation peut être une première étape , mais il faut en général passer à des techniques plus spécialement construites pour le modèle non linéaire.

La variance n'est pas constante

Deux solutions s'offrent à nous:

1) utiliser des transformations de variables pour stabiliser la variance,

2) pondérer.

Ce type de transformation est, en général, considérée dans la situation suivante: on sait que la variable Y suit une loi de probabilité dont la variance est une fonction de la moyenne de cette distribution. Cette connaissance est en général un fait extérieur à l'analyse statistique ; elle est le résultat d'un type de raisonnement comme le suivant:

a) un dénombrement suit une loi de Poisson,

b) pour une variable de Poisson la variance est égale à la moyenne

c) un calcul montre que si on prend la racine carrée du dénombrement, la variance de cette variable transformée est indépendante de la moyenne

d) quand je travaille sur des dénombrements, on prend la racine carrée des

nombres.
Certaines des transformations suivantes ont en plus un avantage: elles rendent plus symétrique les distribution des variables transformées. Elles permettent donc d'être "plus proche" de deux suppositions: la normalité, l'homoscédasticité. Mais, il n'y a ici aucun lien causal ou analytique entre ces deux propriétés: la stabilisation de la variance offre quelquefois en prime la normalisation de la variable transformée.

$Y' = \frac{1}{Y}$	- souvent utilisée quand il s'agit de mesurer la survenue d'un événement dans le temps - stabilisation de la variance - utiliser les moyennes harmoniques
$Y' = \log Y$	égalise les variances quand la variance augmente avec la moyenne - rend additif les effets qui étaient multiplicatifs - normalise souvent la distribution - linéarise une relation y-x dans le cas où la pente augmente avec x en comprimant la partie supérieure de l'échelle
$Y' = Y^2$	- pour les cas opposés à l'utilisation des log: utilisable si la variance diminue quand la moyenne augmente
$Y' = \sqrt{Y}$	- stabilise la variance dans le cas où Y suit une loi de poisson (énumération, dénombrement) - normalise parfois la distribution
$p' = arcsin\sqrt{p}$	- p est une proportion: stabilise la variance
$Y = ln\frac{p}{1-p}$	- p est une proportion - linéarise une relation probit (sigmoide) - ne stabilise pas la variance
$Y = \phi^{-1}(p)$	- p est une proportion, et ϕ est la fonction de répartition d'une loi N(0,1) - linéarise une relation logit (sigmoide) - ne stabilise pas la variance

Transformations fondées sur les observations

On ne possède pas, en général, une connaissance trés approfondie de la distribution des résidus; on peut tout au plus dire que la variabilité des résidus a tendance à augmenter avec une variable quantitative. Cette information empirique ne suffit pas pour améliorer la formulation algébrique du problème ; elle permet cependant de poser un modèle simple: l'écart type des résidus est proportionnel à x, ce qui s'écrit:

$$Var\varepsilon_i = k^2 x_i^2, k > 0$$

Par exemple, si nous considérons le modèle suivant:

$$Y_i = \theta_1 + \theta_2 x_i + \varepsilon_i$$

et si $Var\varepsilon_i = k^2 x_i^2$ alors, on pourra analyser le modèle:

$$Y_i' = \theta_1 x_i' + \theta_2 + \varepsilon_i'$$

avec $Y_i' = \frac{Y_i}{x_i}, x_i' = \frac{1}{x_i}, \varepsilon_i' = \frac{\varepsilon_i}{x_i}$. Dans ce nouveau modèle, $Var\varepsilon_i' = k^2$ et on a conservé la linéarité.

Régression pondérée avec poids a priori

Au paragraphe précédent, nous avons stabilisé la variance en divisant toutes les observations Y_i par x_i quand la variance est proportionnelle à x_i^2. Dans cette méthode, on donne donc un poids faible aux observations connues avec peu de précision, et un poids fort aux observations connues avec peu d'erreur. Il existe une méthode un peu plus générale qui permet, quand on connaît *a priori* la "structure" de la variance, de stabiliser la variance de l'erreur. Elle consiste a affecter un poids ω_i qui doit être déterministe, aux observations. Voyons les conséquences de cette pondération sur un exemple simple: la régression linéaire simple. Considérons donc le modèle

$$Y_i = \theta_1 + \theta_2 x_i + \varepsilon_i.$$

Le critère à optimiser devient

$$SCE(\theta_1, \theta_2) = \sum_{i=1}^{n} \omega_i (Y_i - (\theta_1 + \theta_2 x_i))^2.$$

La seule différence avec la RLS provient du fait que certains termes dans la somme "pèsent" moins que d'autres. En optimisant cette quantité en θ_1 et θ_2, on trouve:

$$\hat{\theta_2} = \frac{\sum_{i=1}^{n} \omega_i (y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{n} \omega_i (x_i - \bar{x})^2}$$

$$\bar{y} = \frac{\sum_{i=1}^{n} \omega_i y_i}{\sum_{i=1}^{n} \omega_i}, \bar{x} = \frac{\sum_{i=1}^{n} \omega_i x_i}{\sum_{i=1}^{n} \omega_i}, \hat{\theta_1} = \bar{y} - \hat{\theta_2}\bar{x}$$

Régression pondérée avec poids fonction des résidus

Quand on souhaite limiter l'influence des points suspects, on fait appel à des techniques dites robustes. Décrivons en quelques mots une approche intuitive: l'idée consiste à donner aux observations un poids d'autant plus grand que les résidus sont faibles. Il faut tout d'abord estimer les paramètres avec des poids *a priori*. De cette première régression, nous obtiendrons des résidus qui pourront nous servir de poids pour la seconde régression. Cette seconde régression permet d'obtenir de nouveaux résidus... Le procédé est itéré jusqu'à ce que les estimations des paramètres ne "bougent" plus d'une itération sur l'autre. [8][9]

[8] Une telle méthode peut aussi être employée quand on pondère par les valeurs ajustées $\hat{Y}$

[9] Il ne faudrait pas croire que comme cette méthode est décrite de façon très intuitive, elle relève du "bidouillage".

Chapitre 3

REGRESSION MULTIPLE

Les idées utilisées en régression multiple sont les mêmes que celles que nous avons vues en régression linéaire simple (RLS). La régression multiple diffère de la RLS par le nombre de variables explicatives présentes dans le modèle. Détaillons un peu la forme générale d'un modèle de régression multiple. Pour cela, nous avons besoin de notations:
comme en RLS, Y sera la variable à expliquer. Les variables explicatives seront au nombre de $p-1$ et seront notées $(x^j)_{j=1..p-1}$. Les paramètres inconnus du modèle seront au nombre de p et seront notés $(\theta_k)_{k=1..p}$. Ces conventions étant faites, on peut maintenant donner la forme générale du modèle:

$$Y = \theta_1 + \theta_2 x_1 + \theta_3 x_2 + ... + \theta_p x_{p-1} + \varepsilon \tag{3.1}$$

Reprenons l'exemple de l'étude des relations entre taille et âge, et ajoutons une nouvelle variable explicative: le poids. Un modèle possible pour expliquer la taille en fonction de l'âge et du poids est le suivant:

$$Y_i = \theta_1 + \theta_2 x_{1,i} + \theta_3 x_{2,i} + \varepsilon_i i = 1..n$$

Y_i est ici la taille du i^{ieme} enfant de l'échantillon
$x_{1,i}$ est l'âge de cet enfant
$x_{2,i}$ représente son poids.
Tout comme en RLS, un terme supplémentaire (ε_i) vient s'ajouter à la fin du modèle qui va traduire le fait que le modèle ne "colle" pas parfaitement aux données observées sur l'échantillon.

Les variables x_1 et x_2 sont supposées déterministes. On voit sur cet exemple, que toutes les variables de la régression multiple, ne jouent pas des rôles symétriques. Ici, la taille (variable à expliquer) est supposée aléatoire, alors que l'âge et le poids sont supposés déterministes. Or, si la taille est aléatoire, il n'existe aucune bonne raison de supposer le poids comme déterministe. L'analyse générale des modèles de régression comportant des variables explicatives aléatoires sort du cadre de ce cours. En revanche, nous reviendrons des modèles particuliers de ce type dans le chapitre consacré à l'analyse de variance.

3.1 Hypothèses

Les hypothèses sous lesquelles fonctionne la régression multiple sont exactement les mêmes que celles que nous avons faites en RLS soit:
1) les variables aléatoires ε_i sont normalement distribuées
2) les $\big(\varepsilon_i\big)_{i=1..n}$ ont une variance (notée σ^2) identique
3) leur espérance est nulle,
4) les $\big(\varepsilon_i\big)_{i=1..n}$ sont indépendants.ł
Les variables x^j étant déterministes, ces trois conditions sont équivalentes aux conditions 1'),2'),3') énoncées pour la RLS.

3.2 Estimation des paramètres

Les paramètres inconnus du modèles $(\theta_1, \theta_2, ..., \theta_p)$ s'estiment exactement en utilisant les mêmes techniques qu'en RLS. En d'autres termes, soit en utilisant la méthode des moindres carrés, soit en utilisant le maximum de vraisemblance. Sous les hypothèses 1,2,3,4 les estimateurs de moindres carrés sont strictement identiques aux estimateurs de maximum de vraisemblance. Ils minimisent la fonction à plusieurs variables suivante :

$$\begin{aligned} SCE(\theta_1, \theta_2, ..., \theta_p) &= \big(Y_1 - (\theta_1 + \theta_2 x_{1,1} + \theta_3 x_{2,1} + ... + \theta_p x_{p-1,1})\big)^2 + ... \\ &\quad \big(Y_2 - (\theta_1 + \theta_2 x_{1,2} + \theta_3 x_{2,2} + ... + \theta_p x_{p-1,2})\big)^2 + ... \\ &\quad + \big(Y_n - (\theta_1 + \theta_2 x_{1,n} + \theta_3 x_{2,n} + ... + \theta_p x_{p-1,n})\big)^2 \\ &= \sum_{i=1}^{n} \big(Y_i - (\theta_1 + \theta_2 x_{1,i} + \theta_3 x_{2,i} + ... + \theta_p x_{p-1,i})\big)^2 \end{aligned}$$

Pour les valeurs estimées que nous noterons $(\hat{\theta_1}, \hat{\theta_2}, ..., \hat{\theta_p})$ la quantité $SCE(\hat{\theta_0}, \hat{\theta_1}, \hat{\theta_2}, ..., \hat{\theta_p})$ est ce que nous avons déjà appelé SCE_{res} c'est à dire, la somme des carrés entre valeurs observées et valeurs ajustés (prévues par le modèle).
Contrairement à la RLS, il est difficile d'exprimer algébriquement sans utiliser le calcul matriciel, les estimateurs des paramètres. Un certain nombre de propriétés peuvent cependant être énoncés sous les hypothèses 1, 2, 3, 4.
Les estimateurs de maximum de vraisemblance de $(\theta_1, \theta_2, ..., \theta_p)$ sont des estimateurs :
- convergents
- sans biais
- de variance minimum
- efficaces
- normalement distribués.

La variance des estimateurs dépend de σ^2 (variance de la population) qui est inconnue a priori. Il faut donc pour avoir une estimation de la variance des estimateurs, et donc pour pouvoir faire des tests d'hypothèses, avoir une estimation de la variance de population. L'estimation (sans biais) de la variance σ^2 est donnée par

$$\hat{\sigma^2} = \frac{SCE_{res}}{n-p} = \frac{1}{n-p}\sum_{i=1}^{n}\left(Y_i - (\theta_1 + \theta_2 x_{1,i} + \theta_3 x_{2,i} + ... + \theta_p x_{p-1,i})\right)^2 = \frac{1}{n-p}\sum_{i=1}^{n}\left(Y_i - \hat{Y_i}\right)^2$$

Avec ces résultats, nous pouvons maintenant faire des tests d'hypothèses.

3.3 Tests d'hypothèses

Nous disposons maintenant des standard error $s.e.$ pour chacun des paramètres soit, pour l'estimateur $\hat{\theta_j}$:$s.e._{\theta_j}$. Différents types de tests peuvent être envisagés: Soit θ_{j_0} un nombre fixé.
Test de $H_0 : \theta_j = \theta_{j_0}$ contre $H_1 : \theta_j \neq \theta_{j_0}$ et pour $k \neq j$, θ_k est quelconque.
On utilise la statistique:$T = \frac{|\hat{\theta_j} - \theta_{j_0}|}{s.e._{\theta_j}}$ et la règle de décision est: si $T > t_{1-\alpha/2}^{n-p}$ alors on rejette l'hypothèse H_0 avec un risque α.
En ce qui concerne les tests d'hypothèses du type
$H_0 : \theta_j = 0$ contre $H_1 : \theta_j \neq 0$ et pour $k \neq j$, θ_k, cette façon de tester (diviser l'estimateur par son s.e.) n'est pas l'unique façon de procéder. Une autre

méthode repose sur la remarque du paragraphe de la RLS. Illustrons la avec notre exemple du début. On veut savoir si le poids à un effet significatif sur la taille. Pour cela, nous allons construire deux modèles: tout d'abord, un modèle dans lequel l'âge et la taille interviennent, ensuite, un modèle avec une "structure" identique au précédent, mais dans lequel nous imposerons au paramètre du poids d'être nul. Pour être un peu plus concret, écrivons ces deux modèles (en respectant les mêmes conventions de notations qu'au chapitre précédent:

$$Y = \theta_1 + \theta_2 x_1 + \theta_3 x_2 + \varepsilon \tag{1}$$

$$Y = \theta_1 + \theta_2 x_1 + \varepsilon \tag{2}$$

Le modèle (1) fournit une SCE résiduelle que nous allons appeler $SCE(\theta_1, \theta_2, \theta_3)$, cette SCE peut s'interpréter comme la part de variation (j'ai aussi employé le mot information) non expliquée par le modèle (1). De même le modèle (2) fournit une SCE résiduelle que nous appellerons $SCE(\theta_1, \theta_2, 0)$ qui peut s'interpréter comme la part de variation non expliquée par le modèle (2).
Par construction, $SCE(\theta_1, \theta_2, 0) \geq SCE(\theta_1, \theta_2, \theta_3)$, en effet, le modèle (1) contient une variable supplémentaire x_2 (le poids) pour expliquer les variations de Y (la taille). Pour un modèle fixé, plus on dispose de données pour calculer une SCE résiduelle, plus cette SCE résiduelle a tendance à augmenter ($n \nearrow \Longrightarrow SCE \nearrow$).
De même, plus le nombre de paramètres estimés dans le modèle augmente, plus la SCE résiduelle diminue ($p \nearrow \Longrightarrow SCE \searrow$). On pourrait très bien ajouter au modèle des variables explicatives qui n'ont rien à voir avec le phénomène étudié et même si ces variables expliquent très peu de dispersion de la variable expliquée, par simple effet mécanique, on observerait une diminution de la SCE résiduelle
Les SCE résiduelles sont donc des quantités qui mesurent les variations non expliquées par le modèle de façon **absolue**.
Les **degrés de liberté** sont les nombres par lesquels il faut diviser les SCE pour les transformer en mesure **relative** des variations (et donc qui rendent comparables les SCE résiduelles estimées dans deux modèles différents) ; on obtient ainsi ce que nous avons appelé en RLS une **variance**. Ainsi le degré

de liberté avec lequel est estimée la SCE résiduelle du modèle général (3.1) vaut $ddl = n - p$, ou encore le nombre de données (n) diminué du nombre (p) de paramètres indépendants estimés dans le modèle.
Nous n'avons évoqué ici, que les degrés de liberté associés à une SCE résiduelle. En fait, chaque fois qu'une SCE est estimée, elle l'est avec un certain nombre de degrés de liberté.[1]

Un instant de réflexion, permet de conclure que la différence $SCE(\theta_1, \theta_2, 0) - SCE(\theta_1, \theta_2, \theta_3)$ que nous noterons $SCE(\theta_3)$ est la part de dispersion expliquée par la variable x_2 (poids). Faisons le bilan des degrés de liberté mis en jeu dans cette opération:
$SCE(\theta_1, \theta_2, \theta_3)$ est estimée avec n-3 degrés de liberté ; n est le nombre de données (30) et 3 correspond aux paramètres estimés dans le modèle $(\theta_1, \theta_2, \theta_3)$.
De même, $SCE(\theta_1, \theta_2)$ est estimée avec n-2 degrés de liberté ; n n'a pas changé de sens, et 2 correspond aux paramètres θ_1, θ_2.
La différence $SCE(\theta_1, \theta_2, 0) - SCE(\theta_1, \theta_2, \theta_3)$ est estimée avec $(n-2) - (n-3) = 1$ degré de liberté (il suffit de faire la différence des degrés de liberté avec lesquels sont estimées les SCE utilisées).

On en déduit que

$$\frac{SCE(\theta_3)}{1} = \frac{SCE(\theta_1, \theta_2, 0) - SCE(\theta_1, \theta_2, \theta_3)}{(n-2) - (n-3)}$$

est la variance expliquée par la variable x_2.łDe même, la quantité $\frac{SCE(\theta_1, \theta_2, \theta_3)}{n-3}$ représente la variance non expliquée par le modèle (2).
Rappelons, de façon un peu intuitive, ce que représente un test de FISHER: un test de FISHER sert à comparer deux variances, il est basé sur une statistique F qui est le quotient de deux variances. La question à laquelle on veut répondre quand on l'utilise pourrait être formulée de la façon suivante: " *la variance du numérateur de F est-elle négligeable devant la variance du dénominateur*".
Transposons cette question à notre exemple: la question *le poids a t-il une influence sur la taille ?* pourrait être transcrite *"les variations de la taille expliquées par la variable poids sont elles négligeables devant les variations*

[1] Une autre définition plus rigoureuse de la notion de degré de liberté sera donnée au chapitre suivant.

non expliquées par les variables prises en compte dans le modèle ?" ou encore, la quantité

$$F = \frac{\frac{SCE(\theta_3)}{1}}{\frac{SCE(\theta_1,\theta_2,\theta_3)}{n-3}} = \frac{SCE(\theta_1,\theta_2,0) - SCE(\theta_1,\theta_2,\theta_3)}{(n-2)-(n-3)} \frac{n-3}{SCE(\theta_1,\theta_2,\theta_3)}$$

est elle "grande ou petite" ?
Si F est "grand" cela signifie que la variance expliquée par la variable poids n'est pas négligeable devant la variance non expliquée par le modèle et, dans ce cas, on dira que l'effet du poids sur la taille est **significatif**. En revanche si F est "petit" alors, la variance expliquée par la variable poids est négligeable devant la variance non expliquée par le modèle, et on dira que le poids n'a pas d'effet significatif [2] sur la taille. Il reste à attribuer un sens à l'expression "F est grand". Cette notion de "grandeur" ne signifie rien dans l'absolu: il va donc falloir se fixer un seuil à partir duquel on dira que F est grand. Si on le fixe très haut, on risque de déclarer à tort que le poids n'a pas d'effet significatif sur la taille ; en revanche, si on le fixe trop bas, on risque de déclarer à tort que le poids à un effet significatif sur la taille. Pour se sortir de ce dilemme, il faut se fixer un risque (le risque de première espèce). Si on note α le risque que l'on se fixe a priori de déclarer à tort que le poids à une influence sur la taille (risque de rejeter H_0 alors que H_0 est vraie) alors le seuil est (pour notre exemple):$f_{1,n-3}^{1-\alpha}$ (ie, la valeur limite au seuil $1-\alpha$ d'une loi de FISHER à 1 et n-3 degrés de liberté).
La méthode que nous venons d'utiliser pour tester l'égalité d'un paramètre à zéro, peut aussi être utilisée pour tester l'égalité de plusieurs paramètres à zéro. Plaçons nous dans le modèle général (3.1) pour décrire la façon de procéder(on supposera que le nombre de paramètres présents dans le modèle est supérieur à 4). Soient par exemple, les hypothèses suivantes à tester:
$H_0 : \theta_2 = \theta_3 = \theta_4 = 0$ contre H_1 : { l'un au moins des paramètres $\theta_2, \theta_3, \theta_4$ est non nul}
Pour tester ces hypothèses, il faut construire deux modèles. Tout d'abord le

[2] On comprend bien ici le sens du mot **significatif**, dire que le poids n'a pas d'effet significatif sur la taille, ne signifie pas que le poids n'a pas d'effet sur la taille, mais plutôt que compte tenu des données dont nous disposons, il est impossible de faire une différence entre l'influence du poids et la variabilité des données

modèle complet (dans lequel tous les paramètres sont présents,

$$Y = \theta_1 + \theta_2 x_1 + \theta_3 x_2 + ... + \theta_p x_{p-1} + \varepsilon \quad (3)$$

ensuite, le modèle complet dans lequel nous avons imposé aux paramètres à tester d'être tous nuls (d'appartenir à H_0).

$$Y = \theta_1 + \theta_5 x_4 + ... + \theta_p x_{p-1} + \varepsilon. \quad (4)$$

Le modèle (3) fournit une SCE résiduelle $SCE(\theta_1, \theta_2, \theta_3, \theta_4, ..., \theta_p)$ estimée avec $n - p$ degrés de liberté, et le modèle (4) fournit une SCE résiduelle $SCE(\theta_1, 0, 0, 0, \theta_5, ..., \theta_p)$ estimée avec $n - (p - 3)$ degrés de liberté.

Par construction, $SCE(\theta_1, 0, 0, 0, \theta_5, ..., \theta_p) > SCE(\theta_1, \theta_2, \theta_3, \theta_4, ..., \theta_p)$ et, comme
$SCE(\theta_1, \theta_2, \theta_3, \theta_4, ..., \theta_p)$ s'interprète comme la part de dispersion non expliquée par le modèle (3) et,
$SCE(\theta_1, 0, 0, 0, \theta_5, ..., \theta_p)$ s'interprète comme la part de dispersion non expliquée par le modèle (4), la quantité :

$$SCE(\theta_1, 0, 0, 0, \theta_5, ..., \theta_p) - SCE(\theta_1, \theta_2, \theta_3, \theta_4, ..., \theta_p) = SCE(\theta_2, \theta_3, \theta_4)$$

s'interprète comme la part de dispersion expliquée par les variables x_1, x_2, x_3. Cette dernière SCE est estimée avec $((n - (p - 3)) - (n - p) = 3$ degrés de liberté. La quantité $\frac{SCE(\theta_2, \theta_3, \theta_4)}{3}$ est donc la variance expliquée par les variables x_1, x_2, x_3.
De plus, $\frac{SCE(\theta_1, \theta_2, \theta_3, \theta_4, ..., \theta_p)}{n-p}$ est la variance non expliquée par l'ensembles des variables présentes dans le modèle (3). On en déduit la statistique de test suivante :

$$F = \frac{SCE(\theta_2, \theta_3, \theta_4)}{3} \frac{n - p}{SCE(\theta_1, \theta_2, \theta_3, \theta_4, ..., \theta_p)}$$

à comparer à$f_{3,n-p}^{1-\alpha}$ (ie, la valeur limite au seuil $1 - \alpha$ d'une loi de FISHER à 3 et n-p degrés de liberté. La règle de décision est la suivante:
si $F > f_{3,n-p}^{1-\alpha}$ alors, on rejette l'hypothèse H_0 avec un risque de première espèce α.ł
Nous verrons un exemple chiffré à la fin du chapitre. Cette façon de procéder sera intensivement utilisée dans le chapitre consacré à l'analyse de variance.

ł

Remarque Les intervalles de confiances de chacun des paramètres estimés se construisent exactement de la même façon qu'en RLS, en d'autres termes, pour un intervalle de confiance de sécurité $1-\alpha$

$$\hat{\theta_j} - t^{n-p}_{1-\alpha/2} s.e._{\theta_j} \leq \hat{\theta_j} + t^{n-p}_{1-\alpha/2} s.e._{\theta_j}$$

3.4 Diagnostic de multicolinéarité

Le terme *colinéarité* provient de l'algèbre linéaire et s'utilise pour dire qu'un vecteur est proportionnel à un autre.
Nous ne traiterons pas ici d'algèbre linéaire, une interprétation plus pratique de ce que représente la colinéarité en statistique pourrait se réduire à un mot : **corrélation**. Pour être plus précis, reprenons notre exemple (relation entre taille et âge, poids).
Supposons qu'il existe une très forte relation entre l'âge et le poids et que les RLS ($taille = \theta_1 + \theta_2 age + \varepsilon$) et ($taille = \theta_1 + \theta_2 poids + \varepsilon$) montrent une très forte liaison entre respectivement (âge et taille) et (poids et taille) (les tests $H_0 : \theta_2 = 0$ pour la première et la seconde régression sont très significatifs).
Que peut-on conclure sur le modèle $taille = \theta_1 + \theta_2 age + \theta_3 poids + \varepsilon$? La réponse est que globalement, l'ensemble des variables (part de variance expliquée par la régression) est significativement lié à la taille. Il est très difficile de connaître la liaison qu'aura individuellement chaque variable explicative (âge et poids) avec la variable à expliquer (taille). Tous les cas de figures sont possibles, par exemple dans ce modèle, il se peut que les tests sur l'effet de l'âge et sur l'effet du poids soient tous les deux non significatifs, ou qu'un seul des deux le soit...
Ce paradoxe peut s'expliquer de la façon suivante: chaque variable explique individuellement une part de dispersion de la taille. Si les variables explicatives sont liées, elles expliquent vraisemblablement une même part de dispersion de la taille. De façon très imagée, on pourrait dire que la première variable (par exemple l'âge) explique toute la variation qu'elle peut expliquer, et, quand la seconde variable "arrive" une grande partie de la variation qu'elle aurait pu expliquer est déjà prise par la première variable (l'âge) ;

cette seconde variable fait donc avec ce qu'elle trouve. En réalité les choses sont un peu plus complexes, "les deux variables devant se partager en même temps la dispersion ".
Pour éviter ce genre de gag, un grand nombre de logiciels fournissent des indices qui donnent des informations sur les liaisons (la colinéarité) entre toutes les variables explicatives. Ces informations sont contenus dans les indices nommés *TOLERANCE* ainsi définis. Pour le paramètre θ_j du modèle général (3.1) la *TOLERANCE* est $1-R^2$, où R est le coefficient de corrélation multiple de la régression entre la variable x_{j-1} et toutes les autres variables explicatives présentent dans le modèle $x_k, k \neq j-1$.

Remarque : Une *TOLERANCE* est un nombre compris entre 0 et 1.
Si elle est égale à 0 cela signifie, que le coefficient de détermination R^2 vaut 1, donc que la variable devant laquelle se trouve le paramètre dont on examine la *TOLERANCE*, est liée de façon déterministe aux autres variables explicatives. On peut en particulier en déduire qu'elle n'apporte aucune information nouvelle sur les variations de la variable à expliquer.
Si la *TOLERANCE* vaut 1, cela signifie, que la variable devant laquelle se trouve le paramètre dont on examine la *TOLERANCE* n'est pas linéairement liée aux autres variables, on peut regarder sans crainte le résultat du test sur le paramètre.

3.5 Ce qu'il ne faudrait pas croire

D'après ce que nous avons vu au chapitre consacré aux tests, un indice qu'il est très important de "contrôler" est la SCE résiduelle (ou encore la variance résiduelle). En effet, tous les tests que nous avons vu en régression multiple sont des tests de FISHER basés sur la statistique F dont le dénominateur est précisément cette variance résiduelle. Le fait de réduire cette variance, rendra tous nos tests plus puissants (dont plus aptes à mettre en évidence des différences).
Nous avons vu que plus le nombre de paramètres (noté p) augmente, plus la SCE résiduelle diminue. On pourrait donc être tenté "d'incorporer" dans le modèle un grand nombre de variables (l'âge du capitaine, sa taille, le nom-

bre de ses maîtresses etc...). Ceci provoquerait une inflation des standard error des paramètres et rendrait in fine les tests moins puissants. De façon imagée, on pourrait dire qu'il y a une certaine part d'information (de dispersion de la variable à expliquer) à partager entre les différents paramètres, plus on met de paramètres dans le modèle, moins chaque paramètre sera porteur d'information. En régression, on est souvent confronté à cette dualité (diminution de la SCE résiduelle vs diminution du nombre de paramètres), il existe des critères qui permettent de choisir un modèle en fonction des objectifs poursuivis (parlez en à votre statisticien habituel), cependant une attitude raisonnée quant au bien fondé de l'utilisation d'une variable en fonction des objectif poursuivis, devrait vous permettre de vous sortir de la plupart des problèmes. Il existe un moyen pour diminuer les standard error des paramètres : augmenter le nombre d'observations. Pour diminuer la variance résiduelle de vos régressions, sans contrepartie, faites appel à votre statisticien préféré qui vous aidera à planifier votre expérience.

3.6 Un exemple

Nous allons étudier les problèmes de l'estimation sur l'exemple que nous avons utilisé depuis le début (relation entre taille et âge, poids). Pour ce faire rappelons les résultats de la première régression que nous avons réalisé: Le modèle utilisé était:

$$Y = \theta_1 + \theta_2 age + \varepsilon \tag{1}$$

et nous avons obtenu:

Regression

Dep var: TAILLE N: 30 Multiple R: .473 Squared Multiple R: .223
Adjusted Squared Multiple R: .196 Standard Error of Estimate: 4.374

Variable	Coefficient	Std Error	Tolerance	T	P(2 tail)
CONSTANT	118.352	3.653	.	32.399	0.000
age	1.406	0.495	.100E+01	2.838	0.008

Analysis of Variance

Source	Sum-of-Squares	DF	Mean-Square	F-Ratio	P
Regression	154.113	1	154.113	8.055	0.008
Residual	535.709	28	19.132		

Pour cette régression, toutes les hypothèses ont été vérifiées.
Si on réalise la seconde RLS :

(2) $$Y = \theta_1 + \theta_3 poids + \varepsilon$$

Regression
Dep var: TAILLE N: 30 Multiple R: .428 Squared Multiple R: .183
Adjusted Squared Multiple R: .154 Standard Error of Estimate: 4.486

Variable	Coefficient	Std Error	Tolerance	T	P(2 tail)
CONSTANT	86.040	16.955	.	5.075	0.000
POIDS	1.521	0.607	.100E+01	2.505	0.018

Analysis of Variance

Source	Sum-of-Squares	DF	Mean-Square	F-Ratio	P
Regression	126.325	1	126.325	6.277	0.018
Residual	563.498	28	20.125		

On en déduit que le poids est significativement lié à la taille :le test du t de Student est significatif à 5% (la vérification des hypothèses ne pose pas de problèmes particuliers).
Passons maintenant à la régression multiple.
Le modèle s'écrit:

(3) $$Y = \theta_1 + \theta_2 age + \theta_3 poids + \varepsilon$$

et on obtient :

Regression

Dep var: TAILLE N: 30 Multiple R: .503 Squared Multiple R: .253

Adjusted Squared Multiple R: .198 Standard Error of Estimate: 4.369

Variable	Coefficient	Std Error	Tolerance	T	P(2 tail)
CONSTANT	99.521	18.562	.	5.362	0.000
age	1.002	0.630	0.6161539	1.590	0.124
POIDS	0.779	0.753	0.6161539	1.035	0.310

Analysis of Variance

Source	Sum-of-Squares	DF	Mean-Square	F-Ratio	P
Regression	174.545	2	87.273	4.573	0.019
Residual	515.277	27	19.084		

Remarquons tout d'abord que d'après les résultats de cette analyse, ni le poids ni l'age n'ont une influence significative sur la taille, ce qui est contradictoire avec les résultats des deux premières analyses. Cependant un examen des résultats un peu plus poussé fait apparaître une *TOLERANCE* de 0.616 (ce qui signifie que la corrélation entre poids et âge vaut:$R = \sqrt{1 - 0.6161539} = 0.62$).
Les variables âge et poids sont donc fortement corrélées, elles apportent vraisemblablement la même information sur la dispersion de la taille.
Analysons maintenant le mécanisme de test que nous avons vu au paragraphe précédent.
Test de $H_0 : \theta_3 = 0$ (le poids n'est pas lié à la taille) contre $H_1 : \theta_3 \neq 0$.
Pour construire la statistique de test, il faut considérer le modèle (3) qui fournit une SCE résiduelle $=SCE(\theta_1, \theta_2, \theta_3) = 515.277$, et le modèle (1) qui fournit une SCE résiduelle $=SCE(\theta_1, \theta_2, 0) = 535.709$.
Nous en déduisons que la part de dispersion expliquée par la variable *poids* est

$$SCE(\theta_1, \theta_2, 0) - SCE(\theta_1, \theta_2, \theta_3) = 535.709 - 515.277 = 20.425$$

et cette quantité est estimée avec $28 - 27 = 1$ degré de liberté. On en déduit que la variance (de la taille) expliquée par la variable poids vaut:$\frac{20.425}{1}$.
Pour construire la statistique F de test sur l'effet de la variable poids sur la taille, il faut diviser la variance expliquée par le poids par la variance résiduelle du modèle (3) qui ici vaut $19.084 = \frac{515.277}{27}$. On obtient alors: $F = \frac{20.425}{19.084} = 1.070$ à comparer à la valeur limite d'une loi de FISHER à 1 et 27. Ou encore en reprenant la relation $T^2 = F$ vue au chapitre de la régression linéaire simple et valable uniquement pour une loi de FISHER à 1

et n degrés de liberté on obtient $\sqrt{F} = \sqrt{1.070} = 1.035$ ce qui est précisément la valeur de T trouvé dans la régression réalisée sur le modèle (3).

Chapitre 4

ANALYSE DE VARIANCE

4.1 Généralités

L'analyse de la variance, est un cas particulier de la régression. La différence essentielle est la "structure" que possèdent les variables explicatives. L'objet de l'analyse de variance est la recherche de relations entre une variable quantitative et des variables qualitatives.
La variable quantitative (notée Y) est la variable à expliquer ; les variables qualitatives aussi appelées **facteurs de variations** (ou plus simplement facteurs) sont les variables explicatives.
Les facteurs sont étudiés à plusieurs **niveaux** .
Par exemple, on veut étudier l'effet de trois pourcentages de concentré dans la ration des porcs sur leur GMQ. La variable à expliquer est $Y = GMQ$, et le **facteur** *pourcentages de concentré dans la ration* (noté R) est étudié sur 3 niveaux:

R_1 10% de concentré
R_2 12% de concentré
R_3 14% de concentré

Dans cet exemple, le but de l'analyse de variance est d'expliquer les variations du GMQ par les différents pourcentages de concentré.
Quand un seul facteur de variation est utilisé pour expliquer les variations de Y, on réalise une analyse de la variance à un facteur (à une voie). Dans le cas général, plusieurs (n) facteurs sont présents pour expliquer les variations

de Y on parle alors d'analyse de la variance à n facteurs [1].

Facteurs étudiés, facteurs contrôlés

Tous les facteurs n'ont pas le même rôle.

Pour s'en convaincre reprenons notre exemple: pour des raisons pratiques il n'est pas possible de faire cette expérience dans un même bâtiment, quatre bâtiments sont nécessaires. Les trois régimes sont distribués dans chaque bâtiment. L'essai est mené, et nous avons maintenant les résultats à analyser. Deux facteurs de variations clairement identifiés sont ici présents: d'une part le régime des porcs, d'autre part, le bâtiment utilisé. Il est possible que le bâtiment utilisé ait une influence sur le GMQ. Ne pas prendre en compte ce facteur risque de laisser une certaine part de variation des GMQ inexpliquée. Notre but n'est pas d'étudier l'effet du bâtiment, mais l'effet du régime. Les facteurs bâtiment et régime n'ont pas le même rôle:
- régime est un **facteur étudié**,
- bâtiment est un **facteur contrôlé**, son unique rôle est d'expliquer certaines variations.

Effets fixes, effets aléatoires

Les deux facteurs que nous avons évoqués dans l'analyse précédente sont fixés, nous avons **choisi** les bâtiments, et les régimes. Les résultats de notre analyse seront donc "extrapolables" à des porcs élevés dans les mêmes conditions (même type de bâtiments etc...)
Dans certains cas, on veut un plus grand degré de généralité dans l'extrapolation des résultats. Par exemple, on veut pouvoir extrapoler les résultats de l'essai à toute une région (un pays...).
Il est bien évident que notre façon de **planifier l'expérience** doit tenir compte de cet objectif.

Il est possible que dans un élevage, le régime 3 soit meilleur que les autres, et que dans un autre élevage, le régime 2 donne des résultats pratiquement

[1] En général, 2 ou 3 facteurs sont pris en compte.

identiques au régime 3. On voit ici que pour pouvoir extrapoler les résultats, il faut prendre en compte dans l'analyse un terme qui exprimera les performances des régimes dans chaque élevage. La question telle qu'elle a été posée, ne peut pas répondre au degré de généralité que nous voulons atteindre.
Une question du type :ł*les effets du régime sont ils (ou pas) négligeables devant les différences de performances de régimes dans les élevages* serait vraisemblablement plus appropriée. Il faut donc, dans notre essai, tenir compte d'un effet élevage.
Passons maintenant à l'aspect pratique de l'essai. Il n'est peut-être pas possible de prendre dans l'essai tous les élevages susceptibles d'utiliser ces doses de concentrés. Il va donc falloir en *"choisir"* un certain nombre. La façon de choisir les élevages est importante. Une bonne méthode consiste à tirer au sort les élevages qui participeront à l'essai. Le facteur *élevage* n'est donc plus un facteur fixé *a priori*, mais un **facteur aléatoire.** En général, on parle plutôt des effets de ce facteur aléatoire en les qualifiant d'**effets aléatoires**.

Degrés de liberté

Le terme de **degré de liberté** est sans doute l'un des plus utilisés en statistique, et c'est certainement celui qui pose le plus de problème quand il s'agit de l'expliquer.
Comme nous l'avons vu, en statistique, on raisonne à partir d'un échantillon pour inférer sur la population. Quand on s'intéresse aux moyennes, le fait de raisonner sur un échantillon ne crée pas de grandes difficultés, car la moyenne d'un échantillon est une estimation de la moyenne de la population. Il est certes peu probable que l'échantillon contienne la plus grande valeur de la variable d'intérêt, mais il est aussi peu probable qu'il contienne la plus petite. L'échantillon ne fournira pas la même moyenne que la population, mais cette moyenne aura autant de chance d'être fausse dans un sens que dans l'autre. La situation est différente quand on considère les mesures de la variation. La possibilité que l'échantillon ne comprenne pas les plus fortes et les plus faibles valeurs, entraîne une sous-estimation de la variation. Cette difficulté est présente chaque fois que l'on utilise pour mesurer la taille d'un échantillon, le nombre de données que contient cet échantillon. Supposons par exemple que nous avons deux données, une seule information suffit à complètement

rendre compte de la dispersion de ces données : c'est la différence[2]. Pour trois observations (notées a, b, c), il suffit de deux données pour décrire parfaitement la dispersion. Par exemple une fois que $a-b$ est connu, la connaissance de $b-c$ ou $a-c$ ou $\frac{a+b}{2}-c$ ou quoi que ce soit d'additif à ce qui est déjà connu , complète la description de la variation. Plus généralement, pour n valeurs, la dispersion peut s'exprimer sous la formes de $n-1$ valeurs. On dit alors que cette dispersion possède $n-1$ degrés de liberté.

4.2 Les modèles

Comme pour la régression, l'analyse de variance repose sur l'écriture de modèles. Certaines formes particulières de modèles seront détaillées dans les paragraphes suivants. Ils seront tous de la forme:

$$Y = \sum effets + \varepsilon$$

La quantité $\sum effets$ pourra être complètement déterministe (modèles à effets fixes) , complètement aléatoire (modèle à effets aléatoires) ou somme d'effets déterministes et d'effets aléatoires (modèles à effets mixtes).
Nous verrons dans chaque paragraphe les hypothèses à faire sur les effets aléatoires.
Comme en régression, ε est appelé **résidu** il mesure les erreurs que nous avons définies dans le chapitre (1). Nous ferons les hypothèses habituelles sur ce terme:
1) la variable aléatoire ε est normalement distribuée
2) la variance des composantes ε est constante
3) les composantes de ε sont indépendantes.ł
Ces hypothèses se vérifient avec les mêmes méthodes qu'en régression (cf chapitre (6)).

[2]rappelons qu'une expression de la variance est donnée par $\frac{1}{2n(n-1)}\sum_{i,j}(X_i - X_j)^2$

Chapitre 5

ANALYSE DE LA VARIANCE A UN FACTEUR

5.1 Modèle à effets fixes

Reprenons l'exemple du début sur des données chiffrées pour illustrer l'emploi de cette méthode.

R_1	R_2	R_3	
500	550	540	
485	560	550	
490	540	540	
495		560	
		550	
492.5	550	548	530

Les marges du tableau représentent, les moyennes du GMQ pour chaque niveau du facteur, et, la moyenne générale.

Prenons des notations:
n est la taille de notre échantillon (ici, n=12)
I est le nombre de niveaux du facteur (ici, I=3= nombre de régimes étudiés)
n_i est le nombre d'observations pour le niveau i du facteur R, dans notre exemple, n_i est le nombre de porc qui ont reçu le régime i
$Y_{i,j}$ est le GMQ du porc numéro j qui a reçu le régime i.
Dans l'exemple, i varie de 1 à 3, et j varie de 1 à n_i, et $n_1 = 4, n_2 = 3, n_3 = 5$.

Le modèle s'écrit:

$$Y_{i,j} = \mu + R_i + \varepsilon_{i,j} \tag{5.1}$$

μ est l'effet moyen général,
R_i est l'effet du niveau i du facteur R
Rappelons que ces quantités $(\mu, (R_i)_i)$ sont des paramètres de population (donc inconnus) il va falloir les estimer.

Estimation des paramètres

μ est estimé par

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^{I} \sum_{j=1}^{n_i} Y_{i,j}$$

En fait, la formule précédente signifie que μ est estimé par la moyenne générale. Ainsi, $\hat{\mu} = \dfrac{500 + 485 + ... + 560 + 550}{12} = 530$
Si $\bar{Y}_i$ est la moyenne du niveau du facteur R ,la quantité R_i (aussi appelée effet différentiel du niveau i du facteur R) est estimée par:

$$\hat{R}_i = \bar{Y}_i - \hat{\mu}.$$

Cette quantité peut être interprétée comme le correctif à apporter à la moyenne générale pour avoir le GMQ des porcs qui ont reçu le régime i. Pour notre exemple, on a: $\hat{R}_1 = 492.5 - 530$, $\hat{R}_2 = 550 - 530$, $\hat{R}_3 = 548 - 530$. Remarquons que par construction, $\sum_{i=1}^{I} \hat{R}_i = 0$ où encore, sur notre exemple:$(492.5 - 530) + (550 - 530) + (548 - 530) = 0$.

Nous pouvons maintenant écrire:

$$Y_{i,j} - \hat{\mu} = \underbrace{\bar{Y}_i - \hat{\mu}}_{R_i} + \underbrace{Y_{i,j} - \bar{Y}_i}_{\varepsilon_{i,j}}$$

En élevant au carré les deux membres de l'égalité précédente, et en en faisant la somme sur toutes les observations, nous obtenons:

$$\sum_{i=1}^{I} \sum_{j=1}^{n_i} (Y_{i,j} - \hat{\mu})^2 = \sum_{i=1}^{I} \sum_{j=1}^{n_i} (\bar{Y}_i - \hat{\mu})^2 + \sum_{i=1}^{I} \sum_{j=1}^{n_i} (Y_{i,j} - \bar{Y}_i)^2$$

En remarquant que

$$\sum_{i=1}^{I}\sum_{j=1}^{n_i}(\bar{Y}_i - \hat{\mu})^2$$

peut se réécrire:

$$\sum_{i=1}^{I} n_i(\bar{Y}_i - \hat{\mu})^2,$$

nous obtenons

$$\sum_{i=1}^{I}\sum_{j=1}^{n_i}(Y_{i,j} - \hat{\mu})^2 = \sum_{i=1}^{I} n_i(\bar{Y}_i - \hat{\mu})^2 + \sum_{i=1}^{I}\sum_{j=1}^{n_i}(Y_{i,j} - \bar{Y}_i)^2 \tag{5.2}$$

où encore:

$$SCE_t = SCE_R + SCE_{Res}.$$

La quantité SCE_t mesure la dispersion totale des GMQ (la quantité totale d'information présente dans les données),
SCE_R mesure la part de dispersion expliquée par le facteur R
SCE_{Res} mesure la part de dispersion non expliquée par le modèle.
A partie de ces SCE, il est maintenant facile de construire les variances correspondantes, il faut avant cela faire le bilan des degrés de liberté:
SCE_t est estimée avec $n-1$ ddl, dans notre exemple la seule donnée de (12-1) nombres suffit à décrire la dispersion
SCE_R est estimée avec $I-1$ ddl. Pour voir ceci il suffit de réécrire SCE_R sous la forme suivante:

$$SCE_R = \sum_{i=1}^{I} n_i R_i^2$$

or comme par construction $\sum_{i=1}^{I} n_i R_i = 0$, si on connaît les $I-1$ premiers R_i le I^{ieme} est aussi connu.
Dans notre exemple SCE_R est estimée avec $3-1$ ddl
Enfin, SCE_{Res} est estimée avec $n-I$ degrés de liberté.
Pour s'en convaincre, il suffit de remarquer que la dispersion pour chaque niveau du facteur est estimée avec $n_i - 1$ ddl et comme le facteur R possède I niveaux, le ddl de la SCE résiduelle est estimée avec $\sum_{i=1}^{I}(n_i - 1) = n - I$.
De façon plus synthétique, **le degré de liberté de SCE_{res} est égal au**

nombre total d'observations n diminué du nombre de paramètres indépendants estimés dans le modèle $1+(I-1)=I$, soit au total $n-I$.
Dans notre exemple, nous disposons de $n=12$ observations, et 4 paramètres ont été estimés dans le modèle :μ, R_1, R_2, R_3. Nous avons vu que par construction les $\hat{R}_i$ ne sont pas indépendants, si on connaît les $I-1$ premiers R_i le I^{ieme} est aussi connu ; seuls $I-1$ paramètres R_i sont indépendants, on en déduit que SCE_{res} est estimée avec $n-I$ degrés de liberté.
Rassurez vous, vous n'aurez pas à faire ces calculs à chaque fois, en général, on utilise la célèbre loi de LAVOISIER: *rien ne se crée, rien ne se perd, tout se transforme...* qui s'applique aussi à l'information :

$$\underbrace{n-1}_{SCE_{tot}} = \underbrace{(I-1)}_{SCE_R} + \underbrace{(n-I)}_{SCE_{res}}$$

Nous pouvons maintenant estimer les variances:
la variance totale est estimée par $V_{tot} = \frac{SCE_{tot}}{n-1}$
la variance expliquée par le facteur R est estimée par $V_R = \frac{SCE_R}{I-1}$
et enfin, la variance résiduelle (non expliquée par le modèle) est estimée par $V_{Res} = \frac{SCE_R}{n-I}$
Grace à l'hypothèse d'homoscédasticité, la variance résiduelle est un estimateur sans biais de la variance de population des $Y_{i,j}$ que nous avons noté σ^2.
On peut donc construire des intervalles de confiance des moyennes $\mu + R_i$ (estimées par $\bar{Y_i}$) en remarquant que $Var\bar{Y_i}$ est estimée sans biais par: $\frac{\hat{\sigma}^2}{n_i}$.
De même, $Var\hat{\mu}$ est estimée sans biais par $\frac{\hat{\sigma}^2}{n}$.
En utilisant ces propriétés, il est alors facile de construire des intervalles de confiance pour les quantités:$R_i - R_j$. Un exemple numérique sera donné plus loin.
Nous disposons maintenant de tous les "ingrédients" nécessaires pour faire des tests.

Tests d'hypothèses

Les hypothèses testées par tous les logiciels de statistique sont les suivantes:
$H_0 : R_1 = R_2 = ... = R_I$ contre
$H_1 : \{$tous les $(R_i)_i$ ne sont pas égaux $\}$
Dans notre exemple, on veut donc savoir si la dispersion des GMQ expliquée

par les régimes peut être confondue (ou pas) avec la dispersion des GMQ non expliquée par le modèle. Cette hypothèse est testée en comparant la variance V_R (qui mesure la dispersion expliquée par les régimes) à V_{res} (qui mesure la dispersion non expliquée par le modèle).
Comme en régression multiple, on utilise le ratio $F = \frac{V_R}{V_{res}}$ qu'il faut comparer à $f^{1-\alpha}_{I-1,n-I}$. Comme d'habitude, $f^{1-\alpha}_{I-1,n-I}$ est la valeur limite au seuil α d'une loi de FISHER à $I-1$ et $n-I$ degrés de liberté.

Résultats de l'analyse

ANOVA

Dep var:GMQ N: 12 Multiple R: .966 Squared Multiple R: .933

Analysis of Variance

Source	Sum-of-Squares	DF	Mean-Square	F-Ratio	P
Regime	8445.000	2	4222.500	62.814	0.000
Error	605.000	9	67.222		

La variable à expliquer est *GMQ*
Nous disposons de *N=12* observations
Le coefficient de détermination (*Squared Multiple R*) mesure le pourcentage de dispersion expliqué par le modèle ici $R^2 = \dfrac{8445.000}{8445.000 + 605.000} = 0.933 = 93.3\%$
La variance expliquée par le facteur *régime* vaut $V_R = \dfrac{8445.000}{2} = 4222.5$, alors que la variance non expliquée par le modèle $V_{res} = \dfrac{605.000}{9} = 67.222$
La statistique de test $F = \dfrac{V_R}{V_{res}} = \dfrac{4222.500}{67.222} = 62.814$. Le test rejette donc l'hypothèse $H_0 : R_1 = R_2 = R_3$ d'égalité des régimes. Le risque de première espèce α de ce test est donné par $P = 0.000 << 5\%$.
Attention, $P = 0.000$ ne signifie pas que le risque α est nul, mais seulement qu'il est très petit.
Il est possible de construire un intervalle de confiance de paramètre de sécurité $1-\alpha$ pour, par exemple, la quantité $R_1 - R_2$.
R_1 est estimé par $\hat{R}_1 = \bar{Y}_1 - \hat{\mu}$,
R_2 est estimé par $\hat{R}_2 = \bar{Y}_2 - \hat{\mu}$,
on en déduit que $R_1 - R_2$ est estimé par

$$(\bar{Y}_1 - \hat{\mu}) - (\bar{Y}_2 - \hat{\mu}) = \bar{Y}_1 - \bar{Y}_2$$

Comme $\bar{Y}_1$ et $\bar{Y}_2$ sont indépendants, la variance de $\bar{Y}_1 - \bar{Y}_2$ est donnée par $Var(\bar{Y}_1) + Var(\bar{Y}_2) = \sigma^2(\frac{1}{n_1} + \frac{1}{n_2})$.
De plus, σ^2 est estimée sans biais par V_{res} avec $n - I$ degrés de liberté, on en déduit donc que :

$$\hat{R}_1 - \hat{R}_2 - t^{n-I}_{1-\alpha/2}\sqrt{\hat{\sigma}^2(\frac{1}{n_1} + \frac{1}{n_2})} \leq R_1 - R_2 \leq \hat{R}_1 - \hat{R}_2 + t^{n-I}_{1-\alpha/2}\sqrt{\hat{\sigma}^2(\frac{1}{n_1} + \frac{1}{n_2})}$$

soit pour un intervalle de confiance à 95% [1]:

$$492.5-550-2.262\sqrt{67.222(\frac{1}{4} + \frac{1}{3})} \leq R_1-R_2 \leq 492.5-550+2.262\sqrt{67.222(\frac{1}{4} + \frac{1}{3})}$$

Remarque : Les mêmes estimations des SCE peuvent être obtenues en utilisant d'abord le modèle:

$$Y_{i,j} = \mu + \varepsilon_{i,j} \tag{5.3}$$

qui fournit une SCE résiduelle qui vaut:9050,
puis le modèle (5.1) qui fournit une SCE résiduelle qui vaut 605.
Par différence, on obtient la SCE expliquée par le facteur *Regime* soit :9050−605. La suite est strictement identique à ce que nous avons vu.

5.2 Modèle à effets aléatoires

Nous avons vu que le modèle à un facteur consistait à considérer I distributions de moyennes $\mu + R_1, \mu + R_2, ..., \mu + R_I$ et à considérer un échantillon de chaque distribution.

$$Y_{i,j} = \mu + a_i + \varepsilon_{i,j} \tag{5.4}$$

Les I paramètres inconnus du modèle (5.4) [2] sont inconnus, toute l'analyse ne concerne que ces I paramètres inconnus et donc seulement les I niveaux du facteur pour lesquels on a fait des observations.

[1] la quantité $t^{n-I}_{1-\alpha/2} = t^9_{0.975} = 2.262$ est la valeur limite au seuil $1 - \alpha/2 = 0.975$ d'une loi de STUDENT à $n - I = 12 - 3$ degrés de liberté

[2] Il n'y a effectivement que I paramètres inconnus:μ et les paramètres $(a_i)_{i=1..I-1}$

On considère maintenant un facteur ayant une "infinité de niveaux possibles" parmi lesquels, on choisit au hasard un échantillon de I niveaux. Pour chacun de ces niveaux, on fait des observations et on veut étudier "l'effet du facteur".
Par exemple, on considère toutes les voitures d'un type donné donc fabriquées de la même façon. Si on prend un **échantillon** de taille I de ces voitures, on peut considérer qu'elles auront des consommations d'essence différentes suivant la voiture. On veut étudier le fait que de façon générale, pour les autos de ce type, la consommation moyenne au 100 km varie d'une voiture à l'autre.
Pour chacune des autos de l'échantillon, on observe la consommation d'essence sur quelques parcours de 100 km on a ainsi des observations $y_{i,j}$ (i est le numéro de l'auto, j est le numéro du parcours).
Pour simplifier le problème, on supposera que toutes les autos effectuent le même nombre de parcours que nous noterons n_0.
Si on prenait le modèle:

$$Y_{i,j} = \mu + a_i + \varepsilon_{i,j} \tag{5.5}$$

où a_i serait déterministe et représenterait l'effet auto, on pourrait simplement comparer entre elles, les I autos de l'échantillon ; ce n'est pas le but recherché.
Il faut en fait considérer, les I réels $a_1, ..., a_I$ comme étant des observations d'une variable aléatoire que nous noterons A dont on veut étudier la distribution. On considère donc que $a_1, ..., a_I$ sont des observations d'une variable aléatoire A dont la distribution est : $\mathbf{N}(0, \sigma_1^2)$.
On a alors le modèle suivant :

$$Y_{i,j} = \mu + A_i + \varepsilon_{i,j} \tag{5.6}$$

où les $(A_i)_{i=1..I}$ sont indépendantes et
les $(\varepsilon_{i,j})_{i=1..I, j=1..n_0}$ sont indépendantes de même loi $\mathbf{N}(0, \sigma^2)$, et A_i et $\varepsilon_{i',j}$ sont
indépendantes pour tout i, i', j. Dans le modèle (5.6),
μ s'interprète comme la consommation théorique de ce type d'auto,
A_i peut s'interpréter comme "l'erreur" de fabrication de l'auto numéro i,
$\varepsilon_{i,j}$ serait "l'erreur" sur le parcours j effectué par l'auto i.
Les paramètres inconnus du modèles sont ici:μ et σ_1^2

Estimation des paramètres

Elle est presque similaire à celles obtenues dans le modèle à effets fixes. Ainsi μ est estimé par:

$$\hat{\mu} = \frac{1}{In_0} \sum_{i=1}^{I} \sum_{j=1}^{n_0} Y_{i,j}$$

c'est la moyenne générale.
En utilisant la même décomposition que dans le modèle à effets fixes,

$$Y_{i,j} - \hat{\mu} = (\bar{Y_i} - \hat{\mu}) + (Y_{i,j} - \bar{Y_i})$$

Nous en déduisons que la variance résiduelle est estimée (comme dans le modèle à effets fixes) par

$$V_{res} = \hat{\sigma}^2 = \frac{1}{I(n_0 - 1)} \sum_{i=1}^{I} \sum_{j=1}^{n_0} (Y_{i,j} - \bar{Y_i})^2$$

Il reste à estimer la variance de l'effet *auto*. Etudions d'un peu plus près la quantité:

$$V_1 = \frac{SCE_A}{n_0(I-1)} = \frac{1}{n_0(I-1)} \sum_{i=1}^{I} \sum_{j=1}^{n_0} (\bar{Y_i} - \hat{\mu})^2 = \frac{1}{I-1} \sum_{i=1}^{I} (\bar{Y_i} - \hat{\mu})^2$$

elle contient la part de dispersion due aux différences entre autos **et** la part de dispersion due aux differences de parcours pour une même auto. Elle n'estime donc pas σ_1^2. En calculant l'espérance de cet estimateur, on obtient :

$$\mathbb{E}(V_1) = \sigma^2 + n_0 \sigma_1^2$$

et il est facile de voir que

$$\mathbb{E}(V_{res}) = \sigma^2$$

On obtiendra un estimateur $\hat{\sigma}_1^2$ de σ_1^2 en calculant la différence :$\frac{(V_1 - V_{res})}{n_0}$. L'estimateur $\hat{\sigma}_1^2 = \frac{V_1 - V_{res}}{n_0}$ est un estimateur sans biais de la quantité σ_1^2, mais cette quantité peut devenir négative (quand $V_1 < V_{res}$). On conviendra donc, quand cette quantité est négative, de la poser égale à zéro. On obtient alors un nouvel estimateur de σ_1^2:

$$\begin{aligned} \hat{\sigma}_1^2 &= (V_1 - V_{res})/n_0 \text{si} V_1 - V_{res} > 0 \\ &= 0 \text{si} V_1 - V_{res} < 0 \end{aligned}$$

Ce nouvel estimateur est malheureusement biaisé.

Tests d'hypothèses

Pour mesurer l'effet du "facteur" auto, nous disposons de $\hat{\sigma}_1^2$.
L'affirmation $\sigma_1^2 = 0$ est équivalente à dire que toutes les autos ont exactement la même consommation sur chacun des n_0 parcours. C'est précisément une question de cette forme que l'on se pose.łConsidérons la question suivante :

on veut savoir si les différences de consommation entre autos sont (ou pas) négligeables devant les différences de consommation entre les parcours. Il faut donc comparer σ_1^2 à σ^2. Par négligeable, on entend

$$\frac{\sigma_1^2}{\sigma^2} < \theta_0$$

Or on sait que la statistique $F = \frac{\hat{\sigma}_1^2}{\hat{\sigma}^2}$ suit sous l'hypothèse $\sigma_1^2 = 0$ une loi de Fisher à $I-1$ et $I(n_0-1)$ degrés de liberté. En utilisant cette remarque, on peut construire une statistique pour tester les hypothèses $H_0 : \frac{\sigma_1^2}{\sigma^2} = \theta_0$ contre $H_1 : \frac{\sigma_1^2}{\sigma^2} > \theta_0$. La règle de décision est la suivante : on rejette H_0 si

$$F > (1 + n_0\theta_0) f^{1-\alpha}_{I-1,I(n_0-1)}.$$

Un exemple

Un essai a été réalisé sur 20 Clio. Chaque Clio a parcourus les 10 mêmes trajets. Une analyse de variance **à effets fixes** à été réalisée avec le modèle

$$Y_{i,j} = \mu + auto_i + \varepsilon_{i,j} i = 1..20, j = 1..10$$

en voici les résultats

Dep var:CONSO N: 200 Multiple R: .640 Squared Multiple R: .410

Analysis of Variance

Source	Sum-of-Squares	DF	Mean-Square	F-Ratio	P
auto	40.527	19	2.133	1.777	?
Error	216.000	180	1.200		

La variance résiduelle est estimée par $\frac{216.000}{180} = 1.200$
on en déduit donc une estimation de σ_1^2: $\frac{2.133 - 1.200}{20} = 0.047$
Par exemple, la statistique de test du test $H_0 : \frac{\sigma_1^2}{\sigma^2} = 0$ contre $H_1 : \frac{\sigma_1^2}{\sigma^2} > 0$ de l'effet du "facteur auto" est obtenue en considérant le rapport: $F = \frac{2.133}{1.200} = 1.777$ à comparer à $(1 + 20 \times 0) f_{19,180}^{0.95} = 33.2$ pour un test avec un risque de première espèce $\alpha = 5\%$.

5.3 Plan et analyse des expériences factorielles

Généralités

Les expériences factorielles permettent à l'expérimentateur d'étudier les effets combinés de deux (ou plus) variables quand elles sont utilisées simultanément.
L'information obtenue d'une expérience factorielle est plus "précise" que celle obtenue d'une série de plans à un facteur, dans le sens où elle permet l'estimation des **interactions**.
La notion d'interaction peut être illustré en utilisant l'exemple suivant:
n étudiants ont passé un examen, et, leur copie a été corrigée par p enseignants (chaque enseignant corrige les n copies). Notons $n_{i,j}$ la note attribuée par l'enseignant j à l'étudiant i.
Si "tout se passe bien", le modèle suivant devrait bien "coller" aux données:

$$n_{i,j} = \mu + E_i + P_j$$

μ représente la note moyenne de la classe E_i est l'effet étudiant, il traduit le fait que tous les étudiants n'ont pas tous la même note P_j est l'effet enseignant, il traduit le fait que tous les enseignants ne notent pas de la même façon.
Supposons maintenant, que le modèle suivant "colle" nettement mieux à la réalité:

$$n_{i,j} = \mu + E_i + P_j + (E * P)_{i,j}$$

les termes μ, E_i, P_j gardent le même sens que précédemment.

Le terme $(E * P)_{i,j}$ est appelé interaction. Il traduit le fait que la façon de noter de l'enseignant j change pour l'étudiant i. Quand une telle interaction existe, la seule connaissance de la façon de noter de l'enseignant j et du niveau de l'étudiant i ne permet pas de prévoir la note $n_{i,j}$.
Quand une interaction existe, le modèle additif ne suffit plus pour décrire les variations de Y.
Planifier une expérience factorielle, c'est avant tout, répondre aux sept questions suivantes:
1) quels facteurs faut il prendre en compte ?
2) combien de niveaux chaque facteur doit il posséder ?
3) pour les variables quantitatives discrétisées, quel espacement choisir entre les niveaux?
4) comment choisir les unités expérimentales?
5) comment contrôler l'erreur expérimentale?
6) quel critère mesure les effets des facteurs?
7) les données expérimentales permettront elles d'estimer les effets des facteurs?
Quand les effets sont fixes, les dimensions d'un **plan d'expérience factoriel** sont données par le nombre de niveaux de chaque facteur pris en compte dans l'analyse.
Par exemple, un plan factoriel à 2 facteurs ayant respectivement 3 et 4 niveaux sera appelé: plan factoriel 3×4.
Dans un plan $3 \times 4 \times 5$, trois facteurs sont pris en compte ; ils ont respectivement 3,4,5 niveaux.
Quand le nombre de niveaux des facteurs est le même, on utilise aussi la notation puissance. Ainsi, un plan à 3 facteurs ayant chacun deux niveaux est appelé plan 2^3.
Un plan d'expérience sera dit **équilibré** si dans chaque **cellule** du plan (chaque croisement de toutes les variables) un même nombre d'observations a été réalisé.
Tous les plans factoriels s'analysent de la même façon, aussi, nous contenterons nous d'un exemple d'analyse d'un plan 2×3.
L'analyse des plans factoriels non équilibrés fera l'objet d'un développement spécial.

5.4 Exemple d'analyse de variance d'un plan factoriel à deux facteurs à effets fixes

On veut étudier l'accroissement du temps d'effet thérapeutique [3] en fonction de trois traitements à deux sites différents.
Un plan factoriel 2×3 équilibré (4 animaux par cellule) a été réalisé, les résultats sont les suivants:

	T_1	T_2	T_3	MOYENNES
SITE 1	4.0	3.6	5.1	
SITE 1	2.3	2.6	6.6	4.4
SITE 1	1.6	2.5	5.9	
SITE 1	6.4	6.0	6.2	
SITE 2	2.4	6.6	5.6	
SITE 2	5.4	6.4	6.4	5.216
SITE 2	3.3	6.9	4.2	
SITE 2	0.8	9.0	5.6	
MOYENNES	3.275	5.45	5.7	4.808

Prenons des notations:
$Y_{i,j,k}$ est la réponse l'animal numéro k ayant reçu le traitement j au site i
k varie de 1 à $n_0 = 4$
j varie de 1 à $J = 3$
i varie de 1 à $I = 2$
Par exemple,$y_{2,3,1} = 5.6$
n est le nombre total d'animaux utilisés dans l'essai, soit:$n = n_0IJ = 4 \times 2 \times 3 = 24$
μ est l'effet moyen général
S_i est l'effet du site i sur la réponse
T_j est l'effet du traitement j sur la réponse
$(S * T)_{i,j}$ est l'effet de l'interaction site i traitement j sur la réponse

[3]Nous utiliserons le terme *réponse* pour évoquer l'effet étudié (ici l'accroissement du temps d'effet thérapeutique)

Cinq modèles linéaires “concurrents” peuvent être écrits:

$$Y_{i,j,k} = \mu + \varepsilon_{i,j,k} \quad (1)$$

$$Y_{i,j,k} = \mu + S_i + \varepsilon_{i,j,k} \quad (2)$$

$$Y_{i,j,k} = \mu + T_j + \varepsilon_{i,j,k} \quad (3)$$

$$Y_{i,j,k} = \mu + S_i + T_j + \varepsilon_{i,j,k} \quad (4)$$

$$Y_{i,j,k} = \mu + S_i + T_j + (S * T)_{i,j} + \varepsilon_{i,j,k} \quad (5)$$

Le terme ε mesure d'une certaine façon l'inadéquation du modèle aux données.
Dans le modèle (1), les effets (s'ils existent) des deux facteurs étudiés sont négligeables devant les effets de tous les facteurs pouvant faire varier la réponse
Le modèle (2) exprime le fait que les variations de la réponse sont expliquées par le site d'administration du traitement, les effets des traitements étant négligeables par ailleurs.

Le modèle (3) exprime le fait que les variations de la réponse sont expliquées par le traitement,les effets du site d'administration étant négligeables.
Le modèle (4) exprime le fait que les variations de la réponse sont expliquées de façon additive par les effets des traitements et les effets du site d'administration. En d'autres termes:
- il existe un traitement “meilleur” que les autres, et ceci quelque soit le site d'administration, et,
- un site d'administration est “meilleur” que l' autre, et ceci quelque soit le traitement utilisé .ł
Enfin, dans le modèle (5), la présence de l'interaction atteste du fait que les effets du traitements dépendent du site d'administration. Ainsi, selon le traitement utilisé, il existe un site d'administration privilégié.
Ces cinq modèles, ne peuvent coexister en même temps, aussi allons nous en “choisir” un. Choisir un modèle implique des conclusions quant aux effets des différents facteurs pris en compte dans l'analyse.
Nous dirons qu'un facteur à un effet **significatif** sur la réponse *si son effet ne peut pas être confondu avec le terme d'erreur* ε. Ceci, rappelons le, ne

signifie nullement que le facteur n'a pas d'effet sur la réponse, mais que cet effet s'il existe, n'est pas suffisant pour être dissocié du *bruit de fond*.
Comme d'habitude, nous ferons les trois hypothèses d'usage sur ε[4]
Dans un premier temps estimons les quantités inconnues des 5 modèles .

Modèle 1

$Y_{i,j,k} = \mu + \varepsilon_{i,j,k}$
Trois quantités sont à calculer dans le modèle (1):
D'une part le paramètre inconnu μ ensuite, une quantité qui mesure la part d'information non expliquée par le modèle (1):SCE_1, et enfin le nombre de degrés de liberté avec lequel SCE_1 est estimé.
μ est l'effet moyen général, il est estimé par la moyenne générale de toute les données:

$$\hat{\mu} = \frac{1}{n_0 IJ} \sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{k=1}^{n_0} y_{i,j,k} = \frac{4.0 + 2.3 + ... + 4.2 + 5.6}{24} = 4.808$$

SCE_1 correspond à la SCE résiduelle du modèle (1).

$$SCE_1 = \sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{k=1}^{n_0} (y_{i,j,k} - \hat{\mu})^2$$
$$= (4.0 - 4.808)^2 + (2.3 - 4.808)^2 + ...(5.6 - 4.808)^2 \approx 95.61833$$

Un seul paramètre est estimé dans ce modèle, donc SCE_1 est estimée avec $ddl_1 = 24 - 1 = 23$ degrés de liberté.

Modèle 2

$Y_{i,j,k} = \mu + S_i + \varepsilon_{i,j,k}$
Il faut dans ce modèle calculer 5 quantités:
- les paramètres inconnus du modèle μ, S_1, S_2
- SCE_2 et ddl_2

Le paramètre μ est l'effet moyen général, son estimation ne diffère pas de celle que nous avons vue pour le modèle (1)
S_i est le "correctif" à apporter à la moyenne générale pour avoir la réponse

[4]Se reporter au chapitre sur la régression

des animaux dont le traitement a été administré au site i.
S_1 est estimé par

$$\begin{aligned}\hat{S}_1 &= \frac{1}{n_0 J}\sum_{j=1}^{J}\sum_{k=1}^{n_0} Y_{1,j,k} - \hat{\mu} \\ &= \frac{4.0 + 2.3 + ... + 5.9 + 6.2}{4 \times 3} - 4.808 \approx 4.4 - 4.808 = -0.408\end{aligned}$$

de même, S_2 est estimé par:

$$\begin{aligned}\hat{S}_2 &= \tfrac{1}{n_0 J}\textstyle\sum_{j=1}^{J}\sum_{k=1}^{n_0} Y_{2,j,k} - \hat{\mu} \\ &= \tfrac{2.4+5.4+...+4.2+5.6}{4\times 3} - 4.808 \\ &\approx 5.216 - 4.808 = 0.408\end{aligned}$$

Par construction, on a la relation $\sum_{i=1}^{I} \hat{S}_i = 0$ ce qui sur notre exemple se résume à :$\hat{S}_1 + \hat{S}_2 = -0.408 + 0.408 = 0$.
On en déduit que SCE_2 vaut :

$$\begin{aligned}SCE_2 = & \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{n_0}(y_{i,j,k} - \hat{\mu} - \hat{S}_i)^2 \\ &= (4.0 - 4.4)^2 + ... + (6.2 - 4.4)^2 + (2.4 - 5.217)^2 + ... + (5.6 - 5.217)^2 \\ &\approx 91.61666\end{aligned}$$

Dans ce modèle, $I + 1 = 2 + 1$ paramètres ont été estimés, cependant, le nombre de paramètres **indépendants** estimés est égal à $I = 2$. En effet, comme il existe une relation entre les $\hat{S}_i$, la seule connaissance des $I - 1$ premiers $\hat{S}_i$ suffit pour en déduire le I^{ieme}. SCE_2 est donc estimée avec $ddl_2 = n - I = 24 - 2 = 22$ degrés de liberté.

Modèle 3

$Y_{i,j,k} = \mu + T_j + \varepsilon_{i,j,k}$
Il faut dans le modèle (3) calculer 6 quantités:
- les paramètres inconnus du modèle μ, T_1, T_2, T_3
- SCE_3 et ddl_3

Le paramètre μ est estimé comme dans le modèle (1)
T_j est le " correctif" à apporter à la moyenne générale pour avoir le temps d'effet
thérapeutique sur des animaux ayant reçu le traitement j.
T_1 est estimé par:

$$\begin{aligned}\hat{T_1} &= \frac{1}{n_0 I}\sum_{i=1}^{I}\sum_{k=1}^{n_0} Y_{i,1,k} - \hat{\mu} \\ &= \frac{4.0 + 2.3 + ... + 3.3 + 0.8}{4 \times 2} - 4.808 \approx 3.275 - 4.808 = -1.534\end{aligned}$$

de même T_2 est estimé par $\hat{T_2} = 5.45 - 4.808 = 0.642$, et T_3 par $\hat{T_3} = 5.7 - 4.808 = 0.892$
Comme pour S_i dans le modèle (2), on a, par construction: $\sum_{j=1}^{J} \hat{T_j} = -1.534 + 0.642 + 0.892 = 0$
On en déduit que SCE_3 vaut:

$$\begin{aligned}SCE_3 = & \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{n_0}(y_{i,j,k} - \hat{\mu} - \hat{T_j})^2 \\ = & (4.0 - 3.275)^2 + ... + (0.8 - 3.275)^2 + (3.6 - 5.45)^2 + ... + (9.0 - 5.45)^2 \\ & + (5.1 - 5.7)^2 + ... + (5.6 - 5.7)^2 \approx 67.155\end{aligned}$$

En tenant le même raisonnement que dans le modèle (2), on obtient : $ddl_3 = n - J = 24 - 3 = 21$.

Modèle 4

$Y_{i,j,k} = \mu + S_i + T_j + \varepsilon_{i,j,k}$
Les estimations des paramètres inconnus de ce modèle sont les mêmes que celles que nous avons obtenues dans les modèles précédents.
Il ne reste plus qu'à calculer la SCE résiduelle et les degrés de liberté correspondants. Pour cela, nous avons besoin des valeurs prédites par le modèle (4). Elle sont regroupées dans le tableau suivant :

	T_1	T_2	T_3
SITE 1	2.87	5.04	5.29
SITE 2	3.68	5.85	6.11

$$
\begin{aligned}
SCE_4 &= \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{n_0}(y_{i,j,k} - \hat{\mu} - \hat{T}_j - \hat{S}_i)^2 \\
&= (4.0 - 2.87)^2 + ... + (6.4 - 2.87)^2 + (3.6 - 5.04)^2 + ... + (6.0 - 5.04)^2 \\
&+ (5.1 - 5.29)^2 + ... + (6.2 - 5.29)^2 + (2.4 - 3.68)^2 + ... + (0.8 - 3.68)^2 \\
&+ (6.6 - 5.85)^2 + ... + (9.0 - 5.85)^2 + (5.6 - 5.6)^2 + ... + (5.6 - 5.6)^2 \\
&\approx 63.15333
\end{aligned}
$$

Les paramètres estimés dans le modèle sont au nombre de $1 + I + J = 6$, mais seulement $1 + (I - 1) + (J - 1) = 4$ sont estimés indépendamment. On en déduit que $ddl_4 = n - 1 - (I - 1) - (J - 1) = 24 - 4 = 20$.

Modèle 5

$Y_{i,j,k} = \mu + S_i + T_j + (S * T)_{i,j} + \varepsilon_{i,j,k}$
Seuls les $I \times J = 2 \times 3 = 6$ paramètres d'interactions sont à estimer dans le modèle 5. Ils mesurent "la distance" qui sépare le modèle (5) de l'additivité des facteurs.
Aussi leurs estimations est donnée par:

$$(S \hat{*} T)_{i,j} = \frac{1}{n_0}\sum_{k=1}^{n_0} Y_{i,j,k} - \hat{Y_{i,j}^4}$$

où $\hat{Y_{i,j}^4}$ est la valeur de Y prédite par le modèle (4) pour la cellule (i, j) (nous les avons calculées, les valeurs prédites sont contenues dans le tableau précédent). Pour estimer les interactions, il faut disposer des moyennes de chaque cellule. Le tableau suivant contient ces moyennes :

	T_1	T_2	T_3
SITE 1	3.575	3.675	5.595
SITE 2	2.975	7.225	5.45

Nous en déduisons:
$(S \hat{*} T)_{1,1} = 3.575 - 2.87$

$(S \hat{*} T)_{1,2} = 3.675 - 5.04$
$(S \hat{*} T)_{1,3} = 5.595 - 5.29$
$(S \hat{*} T)_{2,1} = 2.975 - 3.68$
$(S \hat{*} T)_{2,2} = 7.225 - 5.85$
$(S \hat{*} T)_{2,3} = 5.45 - 5.6$
Il est maintenant facile de calculer SCE_5

$$\begin{aligned} SCE_5 &= \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{n_0}(y_{i,j,k} - \hat{\mu} - \hat{T}_j - \hat{S}_i - (S \hat{*} T)_{i,j})^2 \\ &= (4.0 - 3.575)^2 + ... + (6.4 - 3.575)^2 + (3.6 - 3.675)^2 + ... + (6.0 - 3.675)^2 \\ &+ (5.1 - 5.595)^2 + ... + (6.2 - 5.595)^2 + (2.4 - 2.975)^2 + ... + (0.8 - 2.975)^2 \\ &+ (6.6 - 7.225)^2 + ... + (9.0 - 7.225)^2 + (5.6 - 5.45)^2 + ... + (5.6 - 5.45)^2 \\ &\approx 22.42333 \end{aligned}$$

Par construction, pour tout $j \in \{1, 2, ...J\}$ on a $\sum_{i=1}^{I}(S \hat{*} T)_{i,j} = 0$, de même, pour tout $i \in \{1, 2, ...I\}$,$\sum_{j=1}^{J}(S \hat{*} T)_{i,j} = 0$. Le nombre de termes d'interaction indépendamment estimés est donc:$(I-1)(J-1)$.
On en déduit que le nombre de paramètres indépendants estimés dans le modèle vaut $1 + (I-1) + (J-1) + (I-1)(J-1) = 1 + (2-1) + (3-1) + (2-1)(3-1) = 6$
On en déduit que $ddl_5 = n - \big(1 + (I-1) + (J-1) + (I-1)(J-1)\big) = 24 - 6 = 18$ En utilisant le technique des différences de SCE, on en déduit que: $SCE_1 - SCE_2$ mesure la part d'information expliquée par le facteur S. Cette part d'information est aussi mesurée par $SCE_3 - SCE_4$.
De même, les quantités $SCE_1 - SCE_3$ et $SCE_2 - SCE_4$ mesure la part d'information expliquée par le facteur T.
Comme le plan est équilibré, on a: $SCE_1 - SCE_2 = SCE_3 - SCE_4$ et $SCE_1 - SCE_3 = SCE_2 - SCE_4$
Cette propriété n'est plus vraie quand il s'agit de plan non équilibrés.
La quantité $SCE_4 - SCE_5$ mesure la part d'information expliquée par l'interaction $S * T$
Enfin, rappelons que SCE_5 mesure la part d'information non expliquée par le modèle (5). Cette propriété n'est plus vraie quand il s'agit de plans non

équilibrés. Il est maintenant aisé de calculer les variances: $\frac{SCE_1 - SCE_2}{ddl1 - ddl2} = \frac{SCE_3 - SCE_4}{ddl3 - ddl4} = \frac{400.167}{1} = 4.00167$ est la variance expliquée par le facteur S
$\frac{SCE_1 - SCE_3}{ddl1 - ddl3} = \frac{SCE_2 - SCE_4}{ddl2 - ddl4} = \frac{28.46333}{2} = 14.23167$ est la variance expliquée par le facteur T
$\frac{SCE_4 - SCE_5}{ddl4 - ddl5} = \frac{22.42333}{2} = 11.21167$ est la variance expliquée par l'interaction.
Enfin, $\frac{SCE_5}{ddl_5} = \frac{40.73000}{18} = 2.26278$ est la variance non expliquée par le modèle (5) (variance résiduelle du modèle (5).
Nous pouvons maintenant passer à l'analyse du modèle (5) en regroupant tous ces résultats dans un tableau:

ANOVA

Dep var: Temps N: 24 Multiple R: .758 Squared Multiple R: .574

Analysis of Variance

Source	Sum-of-Squares	DF	Mean-Square	F-Ratio	P
T	28.46333	2	14.23167	6.289	0.008
S	4.00167	1	4.00167	1.768	0.200
T*S	22.42333	2	11.21167	4.955	0.019
Error	40.73000	18	2.26278		

Toutes les statistiques de tests F sont calculées en divisant les variances des effets à tester par la variance non expliquée par le modèle (5) (ou variance résiduelle du modèle 5).
Nous constatons, que l'effet de l'interaction sur la réponse est significatif (à 5 %).
Avant de regarder les tests sur les effets simples (S, et T), il est nécessaire d'analyser cette interaction.
Analyser l'interaction signifie, essayer de comprendre quels sont les effets des facteurs qui "provoque" cette interaction. Généralement, on représente les valeurs de la réponse prédites par le modèle (5) [5] sur un graphique (cf 5.1). Il semble que l'effet des traitements 1 et 3 soit indépendant du site d'administration.
En revanche, pour le traitement 2 l'administration au site 2 donne une augmentation du temps d'effet thérapeutique plus importante qu'au site 1.

[5] ces valeurs sont pour ce modèle particulier les moyennes observées par cellule

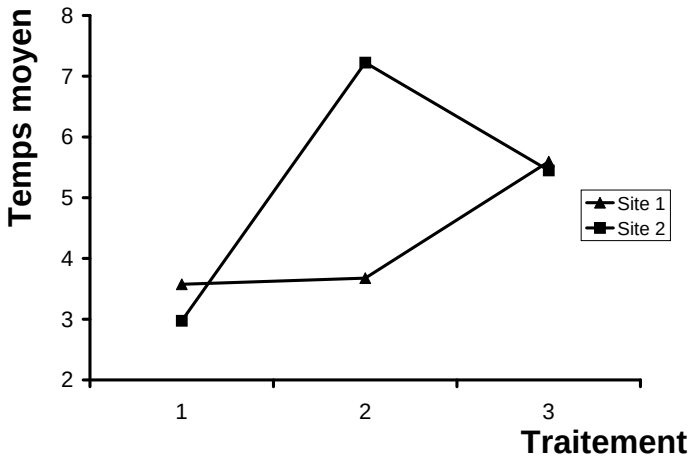

Figure 5.1: Valeurs prédites par le modèle 5. Les courbes ne sont pas parallèles

Que se passerait il si l'interaction était non significative ?
Pour voir les effets d'une interaction nulle, il suffit de se reporter au modèle (4). Le graphique 5.2 montre les valeurs prédites par le modèle (4) en fonction du numéro de traitement, et du site d'administration. On constate que les courbes sont parfaitement parallèles: c'est toujours le cas quand il n'y a pas d'interactions. Aussi, est-il possible de raisonner sur les effets simples.
Si dans le modèle (5) l'interaction était non significative, en d'autres termes si le modèle (4) collait mieux à la réalité, on pourrait dire que les traitements administrés au site 2 augmentent plus la réponse que ceux administrés au site 1. De même, on pourrait dire que le traitement 3 augmente plus que le traitement 2, qui lui même augmente plus que le traitement 1 la réponse.
Si notre objectif est de sélectionner le couple (traitement,site) qui augmente le plus le temps d'effet thérapeutique, on constate que notre conclusion dépendra du modèle choisi.

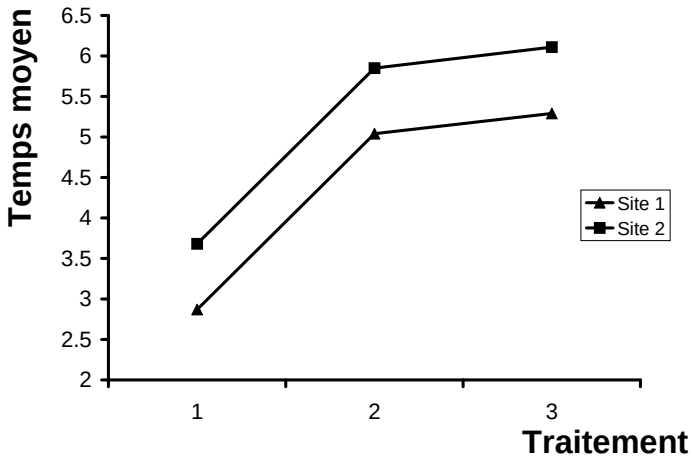

Figure 5.2: Lorsque l'interaction est nulle (modèle 4), les courbes des valeurs prédites sont parallèles

5.5 Analyse d'un plan factoriel non équilibré

Un plan d'expérience factoriel **non équilibré** est un plan dans lequel, toutes les cellules ne contiennent pas le même nombre d'observations.
Un tel plan fait partie de la catégorie des plans **non orthogonaux**.
Avant de donner un sens à ce terme, prenons un exemple.
Un essai sur porcs dont le but est d'évaluer l'action simultanée de trois doses d'antibiotique et de deux doses de vitamines à été mis en place.
Les résultats sont consignés dans le tableau suivant.

	$ANTI_1$	$ANTI_2$	$ANTI_3$
VITA 1	586	559	573
VITA 1	613	573	599
VITA 1	626	599	626
VITA 1	719		
VITA 2	546	479	506
VITA 2	546	506	533
VITA 2	546	546	546
VITA 2	586	546	546

Ce plan n'est pas équilibré, en effet, trois GMQ ont été observés pour la dose 1 de vitamine et pour les doses 2 et 3 d'antibiotique, alors que pour toutes les autres cellules du plan, nous disposons de quatre GMQ.
Ce plan n'est donc pas équilibré.
Pour situer le problème, écrivons les cinq modèles habituels:

$$Y_{i,j,k} = \mu + \varepsilon_{i,j,k} \quad (1)$$

$$Y_{i,j,k} = \mu + VITA_i + \varepsilon_{i,j,k} \quad (2)$$

$$Y_{i,j,k} = \mu + ANTI_j + \varepsilon_{i,j,k} \quad (3)$$

$$Y_{i,j,k} = \mu + VITA_i + ANTI_j + \varepsilon_{i,j,k} \quad (4)$$

$$Y_{i,j,k} = \mu + VITA_i + ANTI_j + (VITA * ANTI)_{i,j} + \varepsilon_{i,j,k} \quad (5)$$

avec les notations suivantes:
$Y_{i,j,k}$ est le GMQ du porc k ayant reçu la dose j d'antibiotique et la dose i de vitamine
j varie de 1 à $J = 3$
i varie de 1 à $I = 2$
k varie de 1 à $n_{i,j}$
$n_{i,j}$ est donc le nombre de GMQ observés dans la cellule (i, j)
n est le nombre total de GMQ observés, ici, $n = 22$
μ est l'effet moyen général
$VITA_i$ est l'effet sur le GMQ de la dose i de vitamine
$ANTI_j$ est l'effet sur le GMQ de la dose j d'antibiotique
$(S * T)_{i,j}$ est l'effet de l'interaction de la dose i de vitamine et de la dose j d'antibiotique sur le GMQ

Les hypothèses d'usage devront être vérifiées. Pour calculer les SCE expliquées par les facteurs $ANTI$, $VITA$, $(ANTI * VITA)$, utilisons la technique des différences de SCE (que nous avons déjà utilisé dans le paragraphe précédent).

En gardant les mêmes notations (SCE_i est la part d'information non expliquée par le modèle i), nous obtenons:
$SCE_1 = 54473.328$ et $ddl_1 = n - 1 = 22 - 1 = 21$
$SCE_2 = 26744.10$ et $ddl_2 = n - I = 22 - 2 = 20$
$SCE_3 = 43857.429$ et $ddl_3 = n - J = 22 - 3 = 19$
$SCE_4 = 18208.767$ et $ddl_4 = n - (I-1) - (J-1) - 1 = 22 - 1 - 2 - 1 = 18$
$SCE_5 = 17740.167$ et $ddl_5 = n - (I-1) - (J-1) - 1 - (I-1)(J-1) = 22 - 1 - 2 - 1 - 2 = 16$
Par différence, on doit obtenir les SCE expliquées par chaque facteur. Intéressons nous dans un premier temps, à la SCE expliquée par le facteur $ANTI$
Elle peut être évaluée en calculant:
$SCE_1 - SCE_3$ ou, $SCE_2 - SCE_4$, soit :
$SCE_1 - SCE_3 = SCE_{ANTI} = 54473.328 - 43857.429 = 10615.899$
$SCE_2 - SCE_4 = SCE_{ANTIajust} = 26744.10 - 18208.767 = 8535.333$
Les SCE ne sont pas égales. De même la SCE expliquée par le facteur $VITA$ peut être évaluée en calculant:
$SCE_1 - SCE_2$ ou $SCE_3 - SCE_4$, soit :
$SCE_1 - SCE_2 = SCE_{VITA} = 54473.328 - 26744.10 = 27729.228$
$SCE_3 - SCE_4 = SCE_{VITAajust} 43857.429 - 18208.767 = 25648.662$
Nous constatons à nouveau que ces SCE ne sont pas égales. Nous nous retrouvons dans la même situation qu'en régression multiple quand les variables sont liées, l'ordre dans lequel les effets des facteurs vont être testés a une influence sur le résultat. En statistique, il existe une interprétation géométrique de cet état de fait.
Quand deux série de données sont liées (corrélées), on dit que les vecteurs ne sont pas orthogonaux, le coefficient de corrélation mesurant le cosinus de l'angle entre ces deux vecteurs. Sous l' hypothèse de normalité des observations, la non corrélation s'interprète en terme d'indépendance, et donc dans ce cas, le mot orthogonalité est synonyme d'indépendance.

De par sa structure, un plan non équilibré est non orthogonal, ceci implique en particulier que les SCE ne sont pas indépendantes.
Pour avoir, pour chaque facteur, des SCE indépendantes il faut tenir compte de la présence de l'autre facteur [6], et par conséquent prendre le modèle dans lequel les deux effets simples sont présents (en d'autres termes le modèle (4)), ainsi par différence, on peut calculer les SCE de chacun des facteurs.
Les degrés de liberté sont encore calculés par différence des degrés de liberté avec lesquels sont estimées les SCE utilisées.
Habituellement, de telles SCE sont dites ajustées.
Le tableau suivant synthétise ces résultats :

SOURCE	SCE	DDL	VARIANCE	F
VITA (compte tenu de ANTI)	25648.662	1	25648.662	23.13***
ANTI	10615.899	2	5307.945	
ANTI (compte tenu de VITA)	8535.333	2	4267.666	3.85*
VITA	27729.228	1	27729.228	
VITA*ANTI	468.600	2	234.3	0.211
RÉSIDUELLE	17740.167	16	1108.760	

où,* est mis pour $P < 0.05$ et *** pour $P < 0.001$.

5.6 Analyse d'un modèle à effets mixtes

Dans les modèles à deux facteurs que nous avons vus, tous les effets étaient fixés. Dans certains cas, l' hypothèse de l'existence d'effets fixes n'est pas raisonnable compte tenu du degré de généralité auquel on veut aboutir.
Un modèle à effets **mixtes** contient des effets **aléatoires**, et des effets **fixes**. C'est le type de modèle le plus couramment rencontré en essai clinique. L'exemple qui suit, illustre la méthode d'analyse de ce type de modèle. Nous nous limiterons ici à donner les différences entre ce modèle et le modèle à effets fixes, à travers un exemple.
Un essai a été mené dans 3 élevages tirés au sort, pour comparer l'effet de différents traitements sur le GMQ de porcs. Dans chaque élevage, chaque

[6] pour un plan à deux facteurs

traitement a été utilisé sur 4 animaux. Les résultats sont les suivants :

	$ELEV_1$	$ELEV_2$	$ELEV_3$
Tt 1	559	586	573
Tt 1	573	613	599
Tt 1	599	626	626
Tt 1	520	719	550
Tt 2	506	479	546
Tt 2	533	506	546
Tt 2	546	546	546
Tt 2	546	546	586

Les 3 élevages présents dans l'essai ont été tirés au sort parmi tous les élevages pouvant utiliser ces traitements. Ils sont donc les représentants de la population des élevages

Ainsi, dans cet essai, un échantillonnage à deux degrés a été réalisé:

- tirage des élevages,
- puis tirage des porcs dans chaque élevage.

Comme les élevages ont été tirés au sort, l'effet élevage est **aléatoire.** Le fait de prendre plusieurs élevages induit donc une dispersion supplémentaire dont il faudra tenir compte dans l'analyse. En revanche, comme nous avons un échantillon d'élevages, les conclusions qui pourront être tirées , pourront [7] être "extrapolées" à la population des élevages (c'est en général le but recherché).

Les différents modèles s'écrivent:

$$Y_{i,j,k} = \mu + \varepsilon_{i,j,k} \quad (1)$$

$$Y_{i,j,k} = \mu + t_i + \varepsilon_{i,j,k} \quad (2)$$

$$Y_{i,j,k} = \mu + E_j + \varepsilon_{i,j,k} \quad (3)$$

$$Y_{i,j,k} = \mu + t_i + E_j + \varepsilon_{i,j,k} \quad (4)$$

$$Y_{i,j,k} = \mu + t_i + E_j + (T * E)_{i,j} + \varepsilon_{i,j,k} \quad (5)$$

avec:

$Y_{i,j,k}$ est le GMQ du porc numéro k ayant reçu le traitement i dans l'élevage

[7] si cet échantillon a été fait en suivant un certain nombre de règles

j
k varie de 1 à $n_0 = 4$
j varie de 1 à $J = 3$
i varie de 1 à $I = 2$
n est le nombre total d'animaux utilisés dans l'essai, soit:$n = n_0IJ = 4 \times 2 \times 3 = 24$
μ est l'effet moyen général
t_i est l'effet du traitement i sur le GMQ
E_j est "l'effet" de l'élevage j sur le GMQ
On supposera que pour tout $j \in \{1, 2, 3\}$ E_j est aléatoire et suit une loi $\mathbf{N}(0, \sigma_1^2)$, et que du fait des conditions d'échantillonnages, les E_j sont indépendantes.
$(T * E)_{i,j}$ est l'effet de l'interaction traitement i dans l'élevage j sur le GMQ.
On supposera que pour tout $(i, j) \in \{1, 2\} \times \{1, 2, 3\}$ $(T*E)_{i,j}$ est aléatoire et suit une loi $\mathbf{N}(0, \sigma_2^2)$, que du fait des conditions d'échantillonnages $(T * E)_{i,j}$ est indépendante de $(T * E)_{i,j'}$ sont indépendantes et que $(T * E)_{i,j}$ et $E_{j'}$ sont indépendantes. Ce terme traduit le fait que dans certains élevages, un traitement est "meilleur" qu'un autre, alors que dans d'autres élevages, ce n'est pas forcément le cas.
Une nouvelle notation est ici utilisée: l'effet traitement est noté en minuscule pour traduire le fait que cet effet est **fixe**, l' effet élevage et l'interaction (traitement*élevage) sont notés en majuscule, pour traduire le fait que ces effets sont **aléatoires.**

Analyse de ce type de plan.

Toutes les variances sont estimées de la même façon que dans un plan à effets fixes. Il faut donc écrire les cinq modèles et calculer par la méthode des différences de SCE les variances de chacun des facteurs.
En revanche, les statistiques de tests utilisées, sont un peu différentes de celles que nous avons utilisées dans le modèle à effets fixes.
Le tableau suivant donne les résultats de l'analyse de variance effectuée en supposant que **tous les effets sont fixes.**

ANOVA

Dep var: GMQ N: 24 Multiple R: .783 Squared Multiple R: .614

Analysis of Variance

Source	Sum-of-Squares	DF	Mean-Square	F-Ratio	P
TRT	21063.375	1	21063.375	17.232	0.001
ELEV	3984.250	2	1992.125	1.630	0.224
TRT*ELEV	9919.750	2	4959.875	4.058	0.035
Error	22002.250	18	1222.347		

Répétons que cette analyse est réalisée en supposant que tous les effets sont fixes.
Par conséquent, les statistiques des tests ne sont pas toutes correctement calculées.

Test de l'interaction:$H_0 : \sigma_2^2 = 0$
La statistique de test à utiliser est exactement la même que pour un modèle à effets fixes par conséquent: $F = \dfrac{4959.875}{1222.347} = 4.058$ convient parfaitement pour tester l'interaction.
Deux cas peuvent se présenter:
- soit l'effet est significatif (ce qui est le cas ici $P < 0.05$) et on continue l'analyse,
- soit l'effet n'est pas significatif, et dans ce cas, il faut recommencer une analyse avec le modèle (4), en d'autres termes en n'incluant pas les interactions dans le modèle.

Test de l'effet élevage:$H_0 : \sigma_1^2 = 0$
Tout comme dans le modèle à effets fixes, on calcule le F correspondant en divisant la variance expliquée par ce facteur par la variance résiduelle. Dans notre exemple, l'effet n'est pas ici significatif (p=0.224).
Test de l'effet traitement:$H_0 : t_1 = t_2$
La statistique ici utilisée n'est pas la bonne.
En effet, on veut répondre à la question:*les effets des traitements sont ils (ou pas)*
négligeables devant les différences de GMQ observées par traitement entre élevages.
Deux cas sont ici possibles:
- soit le test de l'interaction est significatif (ce qui est ici le cas) et le F à utiliser se calcule en divisant la variance expliquée par les traitements par la variance de l'interaction. $F = \dfrac{21063.375}{4959.875} = 4.25$ est à comparer à

$f^{1-\alpha}_{I-1,(I-1)(J-1)}$ qui dans notre exemple pour $\alpha = 0.05$ vaut $f^{0.95}_{1,2} = 200$.
$F = 4.25 << 200$ nous en concluons donc que l'effet des traitements **n'est pas significatif**.
- soit le test de l'interaction n'est pas significatif et dans ce cas, on utilise comme dénominateur de la statistique F la variance résiduelle du modèle (4) avec les degrés de liberté correspondants.[8]
Quand le plan d'expérience est déséquilibré, il faut estimer les variances de chacun des facteurs en tenant compte de ce déséquilibre (se reporter au paragraphe précédent).
L'analyse des modèles à effets aléatoires (tous les effets sont aléatoires) est à peu près identique à celle d'un modèle à effets mixtes.
Les variances de chacun des effets se calculent de la même façon, la seule différence notable réside dans les statistiques de tests à utiliser. Le test de l'interaction ne change pas (on divise la variance de l'interaction par la variance de l'erreur).
En revanche, si le test de l'interaction est significatif, le dénominateur à utiliser pour calculer la statistique de test (F) de tous les effets simples est la variance de l'interaction avec les degrés de liberté correspondants.
Si le test de l'interaction n'est pas significatif, on se reporte alors au modèle additif (l'équivalent de notre modèle (4)) et on utilise la variance résiduelle de ce modèle (avec les degrés de liberté correspondants) pour calculer les statistiques de test des effets simples.

5.7 Une généralisation

Vous avez remarqué qu'il existe un certain nombre de différences entre les plans à effets fixes et les plans à effets mixtes. Ces différences sont de deux ordres :
- conceptuelles tout d'abord, dans un plan à effets mixtes, on suppose que les niveaux du facteurs aléatoires dont nous disposons dans l'expérience sont issus d'un tirage au hasard, alors que dans le plan à effets fixes les niveaux des deux facteurs sont supposés choisis a priori ;
- technique ensuite, les estimateurs des paramètres et par conséquent les tests

[8] Il y a controverse quant à l'attitude à tenir dans ce cas

d'hypothèses réalisés sur les effets des facteurs sont différents dans ces deux types de modèles . Les modèles dans lesquels les deux effets sont aléatoires peuvent aussi être étudiés, et conduisent à des estimateurs des paramètres et à des tests encore différents. Il est possible de présenter de façon unifiée, les différences entre ces types de modèles. Pour simplifier, nous nous bornerons au cas d'un plan à trois facteurs, complet et équilibré.

Soit donc le modèle

$$Y_{i,j,k,l} = \mu + A_i + B_j + C_k + A*B_{i,j} + A*C_{i,k} + B*C_{j,k} + A*B*C_{i,j,k} + \varepsilon_{i,j,k,l}$$

avec

$$i = 1..p,\ j = 1..q,\ k = 1..r,\ l = 1..n.$$

D'un point de vue technique, la différence essentielle entre les trois types de modèle provient du fait que les variances données la machine n'estiment pas les mêmes quantités. Pour élaborer une stratégie générale il faut connaître l'espérance de ces variances. Pour cela nous avons besoin de notations. Nous noterons σ^2_ε la variance non expliquée par le modèle, $\sigma^2_\alpha, \sigma^2_\beta, \sigma^2_\gamma$ sont respectivement les variances expliquées par les facteurs A, B, C. Enfin $\sigma^2_{\alpha\beta}, \sigma^2_{\alpha\gamma}, \sigma^2_{\beta\gamma}$ et $\sigma^2_{\alpha\beta\gamma}$ sont respectivement les variances expliquées par les interactions $A*B, A*C, B*C, A*B*C$. Les quantités $MS_{???}$ représentent les variances données dans le tableau d'analyse de la variance. Par exemple, MS_{AB} est la variance calculée par la machine et "expliquée" par l'interaction AB. Le tableau suivant contient dans la première colonne l'estimateur fournie dans une analyse de variance classique, la deuxième colonne contient l'espérance de cet estimateur quand tous les effets sont fixes, la troisième colonne contient l'espérance de l'estimateur quand tous les effets sont aléatoires, enfin la dernière colonne contient l'espérance de l'estimateur quand le facteur A est aléatoire et les facteurs B et C sont fixes.

Variance	Tous les facteurs fixes	Tous les facteurs aléatoires	A aléatoire B,C fixes
MS_A	$\sigma^2_\varepsilon + nqr\sigma^2_\alpha$	$\sigma^2_\varepsilon + n\sigma^2_{\alpha\beta\gamma} + nq\sigma^2_{\alpha\gamma} + nr\sigma^2_{\alpha\beta} + nqr\sigma^2_\alpha$	$\sigma^2_\varepsilon + nqr\sigma^2_\alpha$
MS_B	$\sigma^2_\varepsilon + npr\sigma^2_\beta$	$\sigma^2_\varepsilon + n\sigma^2_{\alpha\beta\gamma} + np\sigma^2_{\beta\gamma} + nr\sigma^2_{\alpha\beta} + npr\sigma^2_\beta$	$\sigma^2_\varepsilon + nr\sigma^2_{\alpha\beta} + npr\sigma^2_\beta$
MS_C	$\sigma^2_\varepsilon + npq\sigma^2_\gamma$	$\sigma^2_\varepsilon + n\sigma^2_{\alpha\beta\gamma} + nq\sigma^2_{\alpha\gamma} + np\sigma^2_{\beta\gamma} + npq\sigma^2_\gamma$	$\sigma^2_\varepsilon + nq\sigma^2_{\alpha\gamma} + npq\sigma^2_\gamma$
MS_{A*B}	$\sigma^2_\varepsilon + nr\sigma^2_{\alpha\beta}$	$\sigma^2_\varepsilon + n\sigma^2_{\alpha\beta\gamma} + nr\sigma^2_{\alpha\beta}$	$\sigma^2_\varepsilon + nr\sigma^2_{\alpha\beta}$
MS_{A*C}	$\sigma^2_\varepsilon + nq\sigma^2_{\alpha\gamma}$	$\sigma^2_\varepsilon + n\sigma^2_{\alpha\beta\gamma} + nq\sigma^2_{\alpha\gamma}$	$\sigma^2_\varepsilon + nq\sigma^2_{\alpha\gamma}$
MS_{B*C}	$\sigma^2_\varepsilon + np\sigma^2_{\beta\gamma}$	$\sigma^2_\varepsilon + n\sigma^2_{\alpha\beta\gamma} + np\sigma^2_{\beta\gamma}$	$\sigma^2_\varepsilon + n\sigma^2_{\alpha\beta\gamma} + np\sigma^2_{\beta\gamma}$
MS_{A*B*C}	$\sigma^2_\varepsilon + n\sigma^2_{\alpha\beta\gamma}$	$\sigma^2_\varepsilon + n\sigma^2_{\alpha\beta\gamma}$	$\sigma^2_\varepsilon + n\sigma^2_{\alpha\beta\gamma}$
MS_ε	σ^2_ε	σ^2_ε	σ^2_ε

Techniques de tests

Pour comprendre le mécanisme des tests, prenons un exemple. Supposons que l'on veuille tester les hypothèses

$$H_0 : \sigma^2_\alpha = 0 \text{ contre } H_1 : \sigma^2_\alpha > 0$$

Plaçons nous tout d'abord dans le cas où tous les effets sont fixes. on voit que

$$\mathbb{E}(MS_A) = \sigma^2_\varepsilon + nqr\sigma^2_\alpha$$

et que

$$\mathbb{E}(MS_\varepsilon) = \sigma^2_\varepsilon$$

et la quantité

$$F = \frac{MS_A}{MS_\varepsilon}$$

est d'autant plus grande que l'hypothèse H_0 fausse. On peut montrer que si l'hypothèse H_0 est vraie, F est distribué selon une loi de Fisher. On retrouve donc le test classique que les logiciels effectuent. La construction de la statistique de test pour les hypothèses

$$H_0 : \sigma^2_\beta = 0 \text{ contre } H_1 : \sigma^2_\beta > 0$$

est en tout point similaire à celle que nous avons utilisée pour σ^2_α
Plaçons nous tout d'abord dans le cas où le facteur A est aléatoire et les autres fixes. Pour tester les hypothèses

$$H_0 : \sigma^2_\alpha = 0 \text{ contre } H_1 : \sigma^2_\alpha > 0,$$

on utilise la même statistique que dans le cas où tous les facteurs sont fixes (l' espérance de MS_A est identique à celle que nous avons trouvé dans le cas précédent). Il en va tout autrement lorsque on veut tester

$$H_0 : \sigma^2_\beta = 0 \text{ contre } H_1 : \sigma^2_\beta > 0$$

En effet,

$$MS_B = \sigma^2_\varepsilon + nr\sigma^2_{\alpha\beta} + npr\sigma^2_\beta$$

On peut noter que

$$MS_{A*B} = \sigma^2_\varepsilon + nr\sigma^2_{\alpha\beta}$$

Par conséquent le rapport

$$F = \frac{MS_B}{MS_{A*B}}$$

sera d'autant plus grand que l'hypothèse H_0 sera fausse, et on peut montrer que si l'hypothèse nulle est vraie, F est distribué selon une loi de Fisher.
Dans le cas où tous les effets sont aléatoires, on voit qu'il n'existe pas de moyen simple pour tester les effets simples (adressez vous à votre statisticien habituel).

5.8 Plans hiérarchiques à deux facteurs

Généralités

Dans ce type de modèle, un facteur est subordonné à l'autre.
Par exemple, quand on veut comparer l'effet de deux régimes utilisés dans des élevages différents.
Chaque élevage n'utilise qu'un seul régime. Le facteur élevage est alors subordonné au facteur régime, le "choix" des élevages étant fait pour chaque régime, sans qu'il existe de correspondance entre les différents élevages utilisant les différents régimes.
Dans ce cas, comparer la moyenne de l'élevage numéro 1 utilisant le régime 1 à l'élevage numéro 1 utilisant le régime 2 n'a pas de sens.
Voici les résultats d'un essai réalisé dans 2×3 élevages ayant pour but de comparer le GMQ de porcs nourris avec deux régimes différents.

Régimes		1			2	
élevage	1	2	3	1	2	3
porcs	563.6	535.5	555.3	614.5	534.3	585.8
	619.3	597.7	551.1	636.0	549.3	570.2
	558.9	558.9	568.4	646.8	541.5	582.2
	616.9	575.0	520.0	674.9	634.8	613.3
	601.9	565.4	549.9	700.0	626.4	574.4

Des notations:

1. $Y_{i,j,k}$ est l GMQ du porc numéro k ayant reçu le régime i dans l'élevage j
2. k varie de 1 à $n_0 = 5$
3. j varie de 1 à $J = 3$
4. i varie de 1 à $I = 2$
5. μ est l'effet moyen général
6. R_i est l'effet du régime i sur le GMQ
7. $E/R_{i,j}$ est "l'effet" de l'élevage j qui a reçu le régime i sur le GMQ, il est en général appelé: effet élevage dans régime. Cet effet est bien sûr **aléatoire**.

Les modèles hiérarchiques sont donc soit mixtes (l'effet du traitement est considéré comme fixe) soit aléatoire.
Le modèle complet (avec tous les facteurs) s'écrit:

$$Y_{i,j,k} = \mu + R_i + E/R_{i,j} + \varepsilon_{i,j,k}$$

On peut constater que le terme $E/R_{i,j}$ est confondu avec la somme des deux termes $E_j + E * R_{i,j}$ que nous aurions utilisé s'il s'agissait d'un plan factoriel croisé.
Cet effet de confusion n'est d'ailleurs pas seulement algébrique, un instant de réflexion permet de conclure que pratiquement, il n'est pas possible de dissocier ces deux effets.

Première méthode d'analyse

A partir de la remarque précédente, on peut déduire une première méthode d'analyse: On analyse le plan comme s'il s'agissait d'un plan factoriel croisé, en d'autres termes on analyse les données avec le modèle:

$$Y_{i,j,k} = \mu + R_i + E_j + E * R_{i,j} + \varepsilon_{i,j,k}$$

ce qui nous conduit aux résultats suivants:ł

ANOVA

Dep var: GMQ N: 30 Multiple R: .773 Squared Multiple R: .598

Analysis of Variance

Source	Sum-of-Squares	DF	Mean-Square	F-Ratio	P
ELEV	19406.478	2	9703.239	10.610	0.000
REGIME	9958.032	1	9958.032	10.888	0.003
REGIME*ELEV	3321.593	2	1660.796	1.816	0.184
Error	21949.304	24	914.554		

Comme nous savons que les effets E/R et $E + E * R$ sont confondus, on peut calculer la SCE expliquée par E/R en ajoutant les SCE des effets E et $E * R$.

Ce qui nous donne:$SCE_{E/R} = SCE_E + SCE_{E*R} = 19406.478 + 3321.593 = 22728.071$. Cette SCE est calculée avec $J-1+(I-1)(J-1) = 2+2$ degrés de liberté (c'est à dire en faisant la somme des degrés de liberté avec lesquels sont estimées SCE_E et SCE_{E*R}.

Avec ce type d'analyse, on peut répondre à deux questions:

les élevages sont-ils homogènes (pour chaque régime) ?

les régimes sont-ils différents ?

Test sur l'effet du régime

La question posée peut être reformulée de la façon suivante:

la différence entre les régimes est elle supérieure à la différence entre les élevages ?

Les différences entre élevages sont mesurées par la variance de l'effet E/R. La statistique de test à utiliser est donc calculée en faisant le rapport entre la variance expliquée par les régimes et la variance expliquée par E/R (avec les degrés de liberté correspondants).

Test sur l'effet élevage dans régime

Comme pour l'interaction dans le modèle à effets mixte, ce terme se teste en utilisant pour le calcul de F un dénominateur égal à la variance résiduelle.
En résumé, nous avons le tableau d'analyse de la variance suivant :

Dep var: GMQ N: 30 Multiple R: .773 Squared Multiple R: .598

Analysis of Variance

Source	Sum-of-Squares	DF	Mean-Square	F-Ratio
REGIME	9958.032	1	9958.032	1.75
ELEV/REGIME	22728.071	4	5682.018	6.22**
Error	21949.304	24	914.554	

La notation 6.22** signifie que le test est significatif pour $P < 0.01$ Comme pour le modèle à effets mixte, si le test sur $ELEV/REGIME$ n'est pas significatif, on peut réécrire un modèle dans lequel cet effet ne figure plus et tester l'effet du régime par rapport à la variance résiduelle de ce modèle.

Seconde méthode d'analyse

Cette seconde méthode doit être préférée à la méthode précédente dès que le nombre de niveaux du facteur subordonné (au facteur principal), n'est pas le même suivant les niveaux du facteur principal.
Sur notre exemple, on utiliserait cette seconde méthode, si le régime 1 n'avait pas été expérimenté sur le même nombre d'élevages que le régime 2.
Le principe de cette méthode repose sur le réalisation de plusieurs analyses de la variance.
Tout d'abord, il faut réaliser une analyse de variance par niveaux du facteur principal, ici par régime, ce qui nous donne les résultats suivants:

ANOVA 1 sur les 3 élevages utilisant le régime 1

Dep var: GMQ N: 15 Multiple R: .643 Squared Multiple R: .413

Analysis of Variance

Source	Sum-of-Squares	DF	Mean-Square	F-Ratio	P
ELEV	4713.841	2	2356.920	4.227	0.041
Error	6691.200	12	557.600		

ANOVA 2 sur les 3 élevages utilisant le régime 2

Dep var: GMQ N: 15 Multiple R: .736 Squared Multiple R:.541

Analysis of Variance

Source	Sum-of-Squares	DF	Mean-Square	F-Ratio	P
ELEV	18014.230	2	9007.115	7.084	0.009
Error	15258.104	12	1271.509		

Il faut ensuite réaliser une analyse de variance globale (sur tous les élevages) dont voici les résultats:

ANOVA 3 sur les 6 élevages de l'essai

Dep var: GMQ N: 30 Multiple R: .427 Squared Multiple R:.182

Analysis of Variance

Source	Sum-of-Squares	DF	Mean-Square	F-Ratio	P
REGIME	9958.032	1	9958.032	6.241	0.019
Error	44677.375	28	1595.621		

Il suffit enfin de construire le tableau final en additionnant les SCE expliquées par les élevages obtenues dans les analyses de variances 1 et 2, et en répétant cette même opération pour les SCE résiduelles. On obtient :$SCE_{E/R} = 4713.841 + 18014.230 = 22728.071$ et $SCE_{res} = 6691.200 + 15258.104 = 21949.304$.

En n'oubliant pas qu'il faut utiliser le variance de l'effet élevage dans régime pour tester l'effet régime nous obtenons:

ANOVA

Dep var: GMQ N: 30 Multiple R: .773 Squared Multiple R: .598

Analysis of Variance

Source	Sum-of-Squares	DF	Mean-Square	F-Ratio
REGIME	9958.032	1	9958.032	1.75
ELEV/REGIME	22728.071	4	5682.018	6.22**
Error	21949.304	24	914.554	

Ce tableau est exactement identique à celui que nous avons trouvé avec la première méthode.

Cette méthode donne quelques résultats supplémentaires:

- l'analyse 1 nous montre que les élevages qui utilisent le régime 1 ne sont pas identiques ($P < 0.05$)
- l'analyse 2 nous montre la même chose pour les élevages qui utilisent le régime 2 ($P < 0.01$)

On peut, en plus, vérifier que les différences entre élevages, pour chacun des

régimes, sont identiques.
Il suffit de comparer la variance expliquée par les élevages dans la première analyse à celle obtenue dans la deuxième analyse, on obtient: $F = \frac{9007.115}{2356.920} = 3.82$ à comparer à $f_{2,2}^{0.95} = 19$.
Comme $3.82 < 19$, on conclue qu'il n'y a pas de différence significative entre élevages pour un même régime (Il est vraisemblable que le test que nous venons d'effectuer n'est guère puissant).

5.9 Plans en mesures répétées

Ces plans très utilisés en pratiques, sont également dénommés:**SPLIT-PLOT** ou encore **PLANS PARTIELLEMENT HIERARCHIQUES**.

Dans la plupart des expériences, les individus qui ont reçus un même traitement n'ont pas la même réponse à ce traitement.
Cette dispersion **inter-individuelle** rend, en général les tests sur les traitements peu puissants. Si on pouvait séparer cette dispersion "parasite" des dispersions d'intérêts les tests sur les traitements seraient beaucoup plus puissants.
L'un des principaux intérêts des plans en mesures répétés, est de contrôler la dispersion inter-individuelle.
Dans ce type de plans, chaque individu est observé à plusieurs "instants" (dates) (ou sous plusieurs traitements), ainsi, l'effet "instants d'observation" pour l'individu i peut être mesuré par rapport à sa réponse moyenne à tous les instants. Dans un certain sens, chaque individu est utilisé comme son propre contrôle. On pourra alors supprimer la dispersion inter-individuelle pour étudier les dispersion d'intérêt.
Prenons l'exemple de deux groupes d'individus sur lesquelles la pression artérielle est mesurée toute les deux semaines. Le premier groupe est traité avec un médicament anti-hypertenseur standard, le second groupe est traité avec un nouvel anti-hypertenseur. Nous disposons, de plus, pour chaque individu, de sa pression artérielle avant traitement. La première partie du tableau suivant contient les résultats obtenus par les individus traités avec le médicament standard, la seconde contient les résultats des individus traités

avec le nouveau produit.

avt trt	sem 2	sem 4	sem 6	sem 8
102	106	97	86	93
105	103	102	99	101
99	95	96	88	88
105	102	102	98	98
108	108	101	91	102
104	101	97	99	97
106	103	100	97	101
100	97	96	99	93
98	96	97	82	91
106	100	98	96	93
102	99	95	93	93
102	94	97	98	85
98	93	84	87	83
108	110	95	92	88
103	96	99	88	86
101	96	96	93	89
107	107	96	93	97

Une des hypothèses importantes avec laquelle nous allons ici travailler est **l'indépendance** des résultats observés sur un même individu au cours des semaines successives.
Cette hypothèse n'est pas nécessaire dans le cas général, néanmoins l'analyse à laquelle nous serions conduit sans elle est trop compliquée pour être exposée dans ce cours.
Nous verrons à la fin de ce paragraphe, une méthode permettant de vérifier cette hypothèse. Les deux autres hypothèses classiques en analyse de variance devront elles aussi être vérifiées (vous avez tous les outils).
Afin de comparer effectivement les deux traitements, il faut que les individus sont comparables, aussi allons nous travailler sur les différences entre les pressions après traitement et la pression avant traitement.
Nous obtenons alors le tableau suivant:

sem 2	sem 4	sem 6	sem 8
4	-5	-16	-9
-2	-3	-6	-4
-4	-3	-11	-11
-3	-3	-7	-7
0	-7	-17	-6
-3	-7	-5	-7
-3	-6	-9	-5
-3	-4	-1	-7
-2	-1	-16	-7
-6	-8	-10	-13
-3	-7	-9	-9
-8	-5	-4	-17
-5	-14	-11	-15
2	-13	-16	-20
-7	-4	-15	-17
-5	-5	-8	-12
0	-11	-14	-10

L'analyse de ces plans est réalisée avec le modèle général suivant:

$$Y_{i,j,k} = \mu + t_i + s_j + (t * s)_{i,j} + U_{k(i)} + (U * S)_{jk(i)} + \varepsilon_{i,j,k}$$

où:
$Y_{i,j,k}$ est la différence de pression mesurée à la semaine j sur l'individu k soumis au traitement i.
μ est l'effet moyen général
t_i est l'effet du traitement i
s_j est l'effet de la semaine j
$(t * s)_{i,j}$ est l'effet de l'interaction traitement semaine
$U_{k(i)}$ est l'effet de l'individu k subordonné au traitement i (c'est un effet aléatoire). La SCE associé à cet effet sera appelée SCE inter individuelle
$(U * S)_{jk(i)}$ est l'effet de l'interaction individu k soumis au traitement i à la semaine j.
L'effet traitement est dans ce plan **confondu** avec l'effet groupe d'individus, en d'autre termes, il est impossible de faire la différence entre l'effet traitement et l'effet du groupe si les deux groupe ne sont pas homogènes. En revanche, l'effet semaine et l'interaction traitement semaine ne sont confondus avec les effets groupes. Les tests sur les effets semaine et sur l'interaction

traitement semaines seront par conséquent plus puissants que les tests sur l'effet traitement.
Tout comme pour les plans hiérarchiques, l'analyse se décompose en plusieurs analyses.
Tout d'abord, on supprime l'effet traitement en travaillant sur chaque niveau de ce facteur. Il est alors possible de calculer les SCE inter-individuelles par traitement, les SCE intra-traitement, et **l'effet semaine par traitement**. Pour chaque traitement, on utilise donc le modèle suivant:

$$Y_{j,k} = \mu + u_k + s_j + \varepsilon_{j,k}$$

On peut remarquer que dans le modèle ci-dessus, l'effet individu est noté en minuscule, ce qui, avec les conventions que nous avons prises dans le paragraphe consacré aux plans hiérarchiques, signifie que on considère (pour des raisons purement techniques) que l'effet individu est fixe.
La SCE expliquée par l'effet individu est la SCE inter-individuelle.
Pour chaque traitement, la SCE expliquée par l'effet semaine, nous donnera des indications sur l'effet du temps par traitement (en d'autres termes, sur l'interaction temps traitement), nous l'utiliserons un peu plus tard. Enfin, la SCE résiduelle de ce modèle contient les termes d'interaction individu semaine (que nous avons noté dans le modèle complet $(U*S)_{jk(i)}$) [9] et l'erreur expérimentale par traitement, c'est ce que nous avons appelé la dispersion intra-traitement.
Pour notre exemple, nous avons donc deux analyses (une par traitement) dont voici les résultats:

ANOVA POUR LE TRAITEMENT STANDARD

Dep var: DIFF N: 32 Multiple R: .729 Squared Multiple R: .532

Analysis of Variance

Source	Sum-of-Squares	DF	Mean-Square	F-Ratio	P
INDIVIDU	57.500	7	8.214	0.675	0.691
SEM	232.500	3	77.500	6.370	0.003
Error	255.500	21	12.167		

ANOVA POUR LE NOUVEAU TRAITEMENT

[9] pour s'en convaincre il suffit d'écrire le modèle complet soit : $Y_{j,k} = \mu + u_k + s_j + (u * s)_{j,k} + \varepsilon_{j,k}$

Dep var: DIFF N: 36 Multiple R: .772 Squared Multiple R: .596

Analysis of Variance

Source	Sum-of-Squares	DF	Mean-Square	F-Ratio	P
INDIVIDU	114.222	8	14.278	0.840	0.577
SEM	486.972	3	162.324	9.554	0.000
Error	407.778	24	16.991		

Nous pouvons, à ce stade de l'analyse, vérifier l' hypothèse d'homogénéité des groupes en testant l'égalité des variances individus par traitement, ce qui ici donne: $F = \frac{14.278}{8.214} = 1.74 < f_{8,7}^{0.95} = 3.73$.
Comme le F calculé est inférieur à la valeur limite trouvée dans la table, le test ne contredit pas l'hypothèse d'homogénéité des groupes. On peut cependant noter, la puissance relativement faible de ce type de test: les variances étant estimées avec très peu (8 et 7) de degrés de liberté.
Il est aussi possible de comparer les variances résiduelles de ces modèles, qui mesurent, en plus de l'inadéquation du modèle aux données, le comportement "chaotique" que peut avoir un individu donné soumis à un traitement donné. Le test sur les variances résiduelles (variances intra) conduit à :$F = \frac{16.991}{12.167} = 1.31 < f_{24,21}^{0.95} \approx 2.12$.
A nouveau, le test ne contredit pas l'hypothèse d'homogénéité des groupes.
La somme des SCE expliquées par le facteur individu pour chaque traitement, est donc la SCE globale expliquée par l'effet individu, ou encore la SCE inter-individuelle: elle vaut ici: $57.5 + 114.222 = 171.722$
De même, la sommes des SCE résiduelles obtenues pour chaque traitement, est donc la SCE intra-traitement globale, elle vaut ici:$255.5 + 407.778 = 663.278$
Enfin, la somme des SCE expliquée par le facteur semaine est l'effet semaine **compte-tenu** de l'effet traitement.Elle vaut ici:$232.5 + 486.972 = 719.472$
Nous disposons de presque toutes les informations sur les effets individus, il reste à étudier les effets traitements, semaine, et surtout l'interaction entre ces deux effets.
Voyons donc, la seconde partie de l'analyse.
Le modèle suivant, construit sur l'ensemble des individus, va nous servir à estimer l'effet traitement, l'effet **simple** semaine, et l'interaction entre ces deux effets:

$$Y_{i,j,k} = \mu + t_i + s_j + \varepsilon_{j,k}$$

Ce modèle nous conduit aux résultats suivants:

ANOVA SUR L'ENSEMBLE DES DONNEES

Dep var: DIFF N: 68 Multiple R: .703 Squared Multiple R: .495

Analysis of Variance

Source	Sum-of-Squares	DF	Mean-Square	F-Ratio	P
TRT	196.160	1	196.160	13.967	0.000
SEM	669.691	3	223.230	15.895	0.000
Error	884.781	63	14.044		

La SCE expliquée par le facteur semaine est ici calculée sur l'ensemble des individus, sans tenir compte de l'effet du traitement, par différence avec celle que nous avons calculée en tenant compte du facteur traitement, nous obtiendrons la SCE expliquée par l'interaction traitement*semaine, soit :$719.472 - 669.691 = 49.781$.

Comme dans un plan factoriel classique, cette interaction est estimée avec : $(nbsem - 1) \times (nbtrt - 1) = 3 \times 1 = 3$ degrés de liberté. $nbsem$ et $nbtrt$ représentent respectivement le nombre de semaines et le nombre de traitements.

Il ne reste plus qu'à récapituler tous ces résultats dans un tableau :

Source	Sum-of-Squares	DF	Mean-Square	F-Ratio
INDIVIDU	171.722	15	11.448	
SEM	669.691	3	223.230	15.14***
TRT	196.160	1	196.160	17.13***
SEM*TRT	49.781	3	16.593	1.1
Error(intra)	663.278	45	14.74	

*** signifie que pour ce test $P < 0.001$.

Le test réalisé sur l'effet semaine est construit en divisant la variance expliquée par l'effet semaine par la variance résiduelle (intra traitement):$F = \frac{223.230}{14.74} = 15.14 > f_{3,45}^{0.999} \approx 6.6$

Le test réalisé sur l'effet traitement, répond à la question:

la différence d'effet entre les traitements est-elle supérieure à la différence entre individus,

ainsi, la statistique de test F est obtenue en calculant le rapport entre les variances expliquées par le facteur traitement et celle expliquée par le facteur individu, soit $F = \frac{196.16}{11.448} = 17.13 > f_{1,15}^{0.999} = 4.54$
Le test sur l'interaction est obtenu en calculant le rapport entre la variance expliquée par l'interaction et la variance résiduelle soit: $F = \frac{16.593}{14.74} = 1.1 < f_{3,45}^{0.95} \approx 2.80$
Conclusions: Les traitements sont significativement différents, et la pression artérielle évolue aux cours du temps.
Il est ici possible d'aboutir à une conclusion car l'interaction semaine traitement est non significative. Si cette interaction était significative, il faudrait représenter graphiquement l'évolution des pressions artérielles moyennes au cours du temps pour les deux traitements.
Le fait que l'interaction soit significative s'expliquant par un non parallélisme des courbes.
Deux cas de figures se présentent alors:
- soit les courbes se coupent, et dans ce cas les traitements ne sont pas comparables sur la totalité des temps d'observations,
- soit les courbes ne se coupent pas et dans ce cas on peut conclure.ł
Il faut maintenant vérifier l'indépendance des observations.
Il faut calculer les résidus et calculer les corrélations entre les résidus obtenus pour les résultats d'une semaine donnée avec les autres semaines. Ces corrélations sont consignées dans la tableau suivant :

	SEM2	SEM4	SEM6	SEM8
SEM2	1.000			
SEM4	-0.375	1.000		
SEM6	-0.639	0.158	1.000	
SEM8	0.075	0.307	0.067	1.000

On voit que les corrélations entre les résidus de toutes les semaines ne sont pas très fortes:la plus forte est inférieure (en valeur absolue) à 0.639.
Aucune structure particulière ne se dégage de ce tableau.

Conclusion

Tous les thèmes présentés dans ce cours ne sont pas classiquement présentés dans les cours d'introduction au modèle linéaire. J'ai choisi de présenter ce qui me parait le plus utile pour l'utilisation rapide de cette famille de modèle dans un contexte vétérinaire. Un grand nombre de classes de modèles (modèles mixtes, séries chronologiques...) mériteraient une étude approfondie qui dépasse les ambitions de cette introduction. Il va de soi que cette initiation n'est nullement suffisante pour qui veut/doit comprendre les tenants et aboutissants des méthodes présentées ici. Dans cette situation, je vous engage à lire l'abondante littérature consacrée à ce modèle.

Vous pouvez par exemple commencer par les ouvrages suivants faciles et agréables à lire :

Dagnelie P. [1998]. Statistique théorique et appliquée. Tome 1. Statistique descriptive et bases de l'inférence statistique. Paris et Bruxelles, De Boeck et Larcier, 508 p.

Dagnelie P. [1998]. Statistique théorique et appliquée. Tome 2. Inférence statistique à une et à deux dimensions. Paris et Bruxelles, De Boeck et Larcier, 659 p.

Draper N. R., and Smith H. [1981] Applied Regression Analysis. 2nd ed., New York, John Wiley and Sons, 709 p.