Statistique descriptive 2
RES 2-05

Paul-Marie Grollemund

Table des matières

CHAPITRE 1

Introduction

Le propos de cette ressource pédagogique est d'apporter des notions et des outils utiles à la détection et à la caractérisation de liaisons entre deux facteurs quantitatifs. L'enjeu autour de ces types de liaisons est de pouvoir en déduire une information concernant une variable, en fonction de ce qu'on connait sur une autre variable. Pour illustrer cela, nous utilisons dans ce chapitre l'exemple d'une étude pour laquelle les tailles de pères et de fils sont analysées, de sorte à évaluer si la taille du père peut être un facteur déterminant pour caractériser les variations de tailles qui sont observées sur un ensemble d'enfants.

Pour étudier ce type de liaisons, il est plus communément connu d'utiliser des approches linéaires, mais par opposition des approches non-linéaires peuvent aussi être pertinentes suivant le contexte. Linéaire et non-linéaire sont des termes qui délimitent les approches possibles et les chapitres suivants détaillent leurs différences.

La suite de ce document se structure en trois chapitres, et le chapitre 2 donne une présentation générale de ce en quoi consiste une liaison ou une absence de liaison entre deux variables quantitatives. Le chapitre 3 aborde les approches linéaires en détaillant la modélisation sous-jacente à cette approche, ainsi que les méthodes de calculs à mettre en œuvre afin d'obtenir en pratique les résultats numériques. Le chapitre 4 donne une introduction de l'approche non-linéaire en introduisant quelques modèles standards.

Table des matières de ce chapitre

1.1 Rappel d'analyse descriptive bivariée

Dans le cadre de la ressource pédagogique "Statistique descriptive 1", des outils d'analyse bivariée ont été introduit. Ces outils couvrent les traitements possibles à effectuer suivant la nature de la paire de variables à étudier. Par exemple, si les deux variables sont qualitatives, il faut utiliser le tableau de contingence (et ses variantes distributionnelles, comme les profils lignes et les profils colonnes) pour décrire et étudier la liaison entre ces deux variables. Si les deux variables sont des variables quantitatives, il est possible de les représenter conjointement à l'aide d'un nuage de points, voir la figure 1.1 pour un exemple, ou de calculer le coefficient de corrélation linéaire empirique dont l'expression est :

$$r(x,y) = \frac{\text{cov}(x,y)}{s_x s_y}$$

où les termes de ce ratio sont :

$$\text{La covariance :} \quad \text{cov}(x,y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})$$

$$\text{L'écart-type :} \quad s_x = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

La valeur du coefficient de corrélation linéaire est dans l'intervalle $[-1,1]$ et voici un guide d'interprétation (qui n'est pas absolu et qui reste à l'appréciation subjective et à la particularité du contexte étudié) concernant la liaison potentielle entre les deux séries de données x et y :

- si $r \in]0.6,1]$, on peut en déduire qu'il y a une forte liaison linéaire positive,
- si $r \in]0.2,6]$, on peut en déduire qu'il y a une faible liaison linéaire positive,
- si $r \in [-0.2,0.2]$, on peut en déduire qu'il n'y a pas une liaison linéaire,
- si $r \in [-0.6,-0.2[$, on peut en déduire qu'il y a une faible liaison linéaire négative, et

- si $r \in [-1, -0.6[$, on peut en déduire qu'il y a une forte liaison linéaire négative.

Remarque 1.1.1 (Liaison positive ou négative)**.** *Pour avoir le détails de ce qu'il faut comprendre concernant la terminologie "liaison positive" ou "liaison négative", il faut consulter le chapitre 2.*

La corrélation linéaire et le nuage de points permettent de n'apporter qu'une analyse descriptive limitée concernant la liaison entre deux variables quantitatives. Dans le cadre de cette ressource pédagogique, de nouveaux outils sont introduits afin de compléter l'analyse qui serait faite seulement avec ces deux outils. Les objectifs complémentaires sont de détailler plus finement le type de liaison, mais aussi de pouvoir tirer profit de la caractérisation de cette liaison pour calculer des prévisions.

1.2 La régression : du père au fils

Pour illustrer le principe de l'approche de régression qui est détaillé dans le reste de ce document, nous étudions ici des données collectées par Gatson en 1885, et dont les conclusions de l'étude ont donné son nom à la "régression". Les données étudiées sont une collection d'informations concernant 205 familles. On dispose des mesures de tailles à l'âge adulte des deux parents et des enfants, et l'étude de Gatson s'intéresse en particulier à une association contre-intuitive entre la taille du père et la taille des fils. Le constat est le suivant ; les pères dont la taille est au-dessus de la moyenne ont tendance à avoir des fils plus petits qu'eux, et inversément. Autrement dit, il y a une "régression" (une baisse) de la taille des fils par rapport à la taille des pères, pour ce qui est des pères assez grands.

Pour illustrer cela, la figure 1.1 est un nuage de points pour lequel on constate effectivement cette "régression". Le complément apportait par ce graphique est aussi la droite (qu'on appelle droite de régression), qui modélise quelle est la différence moyenne de taille en fonction de la taille du père. Cela n'indique pas que les différences de taille sont exactement calculables à l'aide de l'équation de cette droite, mais que mises de côté des variations individuelles, la taille du fils est en moyenne autour de la valeur obtenue par l'équation de la droite de régression. Pour le reformuler plus grossièrement, la régression de Gatson consiste à déterminer la droite qui synthétise le

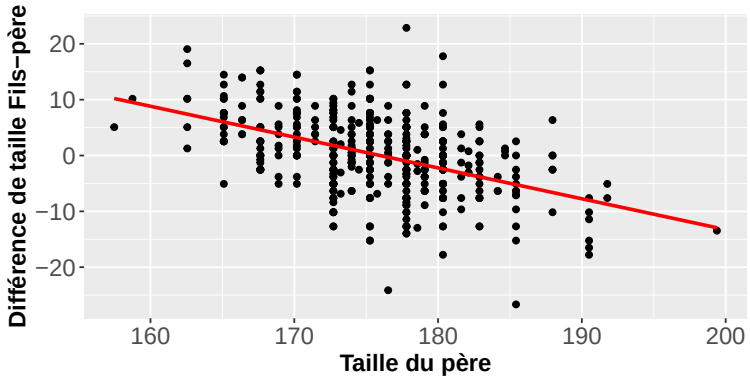


Figure 1.1 – *Nuage de points des données de Gatson concernant les tailles respectives des pères et de leurs fils.*

mieux l'évolution (de gauche à droite) du nuage de points, à savoir la droite qui "épouse" le mieux possible ce nuage de points.

1.3 Les objectifs de cette ressource pédagogique

Dans le cadre de cette ressource pédagogique et de ce document, les objectifs sont de connaître les notions relatives aux différentes approches de régression, ainsi que savoir les mettre en œuvre. En particulier, au terme de cette ressource, il est nécessaire de maitriser et mettre en pratique les notions suivantes :

- une liaison statistique,
- qu'est-ce qu'une liaison linéaire ou l'absence de liaison linéaire,
- la régression linéaire simple et calculer des prévisions,
- la méthode d'ajustement et évaluer la qualité d'ajustement, et
- les régression non-linéaires et faire un choix parmi plusieurs modèles possibles.

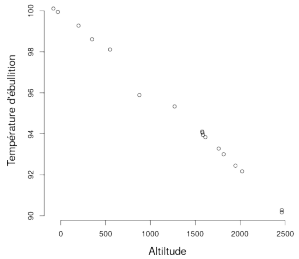
1.4 Diapos de cours

Chap. 1 – **Introduction**

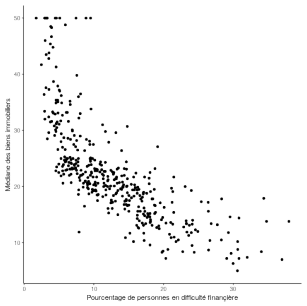
Statistique descriptive : ajuster un modèle

Contenu de cette ressource pédagogique :
- Etablir et quantifier un lien entre deux variables quantitatives
- Ajuster une liaison linéaire
- Ajuster une courbe (liaison non-linéaire)
- Prévision et prédiction de nouvelles valeurs

Définition : Ajustement
Déterminer une estimation des paramètres du modèle à partir des observations. Le modèle obtenu est alors **ajusté** aux données.



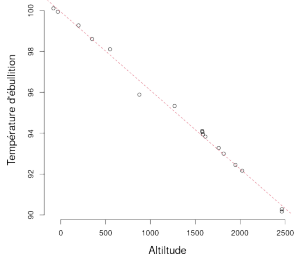
Définition : Ajustement de courbes
Calibrer un modèle pour qu'il *épouse* la forme d'un nuage de points. Cela revient à ajuster un modèle qui prédit une des deux variables à partir de l'autre.



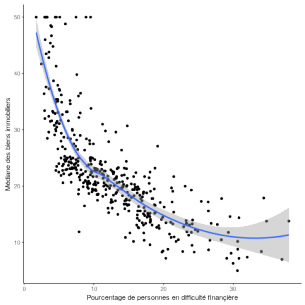
Statistique descriptive : ajuster un modèle

Contenu de cette ressource pédagogique :
- Etablir et quantifier un lien entre deux variables quantitatives
- Ajuster une liaison linéaire
- Ajuster une courbe (liaison non-linéaire)
- Prévision et prédiction de nouvelles valeurs

Définition : Ajustement
Déterminer une estimation des paramètres du modèle à partir des observations. Le modèle obtenu est alors **ajusté** aux données.



Définition : Ajustement de courbes
Calibrer un modèle pour qu'il *épouse* la forme d'un nuage de points. Cela revient à ajuster un modèle qui prédit une des deux variables à partir de l'autre.



Ressource : RES 2-05

Objectif du module :
Connaître la théorie de l'ajustement (méthode des moindres carrés)
Savoir effectuer un ajustement linéaire
Savoir effectuer un ajustement non-linéaire, pour quelques modèles standards
Evaluer la qualité de l'ajustement (R^2)
Savoir prédire de nouvelles valeurs

Enseignant à contacter : en cas de problèmes ou de questions Mail à paul_ marie (dot) grollemund (at) uca (dot) fr
Ou se rendre au bureau des enseignants STID.

Déroulement de cette ressource pédagogique

TD : 5 séances
Semaines 10 à 14
Enseignant pour chacun des groupes : Paul-Marie Grollemund

TP : 5 séances (SAE "Régression sur données réelles")
Semaines 12 à 15 et semaine 18
Enseignant pour chacun des groupes : Paul-Marie Grollemund

Contrôle :
Moyenne des contrôle continu (en fin de séances de TD).
Contrôle de connaissance pendant la semaine 19.
Contrôle d'exercice pendant la semaine 19.

Note finale : moyenne des 3 notes.

CHAPITRE 2

Liaisons statistiques

Au préalable des notions de régression, il est nécessaire d'introduire ce qu'il faut entendre par la terminologie de "liaison statistique". Dans ce qui suit, ce chapitre se décompose en des parties traitant du contexte global (la section 2.1), de ce qu'est une liaison statistique (la section 2.2), des types de liaisons (linéaire en section section 2.3, et non-linéaire en section 2.4), et de ce que veut dire "causalité" (la section 2.5).

Table des matières de ce chapitre

2.1 Contexte et notations

Pour un ensemble de n unités statistiques, on dispose de deux séries de données quantitatives continues. On note les $x = (x_1, \ldots, x_n)$ et $y = (y_1, \ldots y_n)$. De plus, il est à noter que dans le contexte de ce document, une paire de données (x_i, y_i) est relative à une seule et même unité statistique. Autrement dit, pour le $i^{\text{ème}}$ individu, on mesure deux caractéristiques différentes, l'une qui est notée x et l'autre qui est notée y. Bien qu'on définisse ici les données qu'on obtient en pratique, les notions introduites dans la suite de ce chapitre sont principalement des notions théoriques. L'objectif est d'introduire au préalable les notions théoriques, à savoir les quantités qui sont inconnues en pratique, et ensuite de définir dans les chapitres 3 et 4 les approches pratiques et les quantités empiriques nécessaires pour estimer/approcher les quantités théoriques et inconnues.

Les données (x_i, y_i), pour $i = 1, \ldots, n$, sont des mesures associées à des variables aléatoires X et Y, qui correspondent à des phénomènes qu'on souhaite étudier. Dans un contexte concret, il est généralement commun qu'on souhaite particulièrement étudier un seul de ces deux phénomènes. Le deuxième est alors étudié dans l'espoir qu'il y ait une relation entre ces deux phénomènes, et que cette relation puisse nous permettre d'étudier plus finement le phénomène d'intérêt.

Définition 2.1.1 (Variable à expliquer) *Correspond à la variable qu'on souhaite principalement étudier, parmi l'ensembe des variables mesurées.*

Définition 2.1.2 (Variable explicative) *Correspond à une variable à utiliser pour étudier une "variable à expliquer".*

Notation 2.1.3 (Variables explicative et à expliquer)**.** *Par convention, on note Y la variable qu'on souhaite expliquer, et X la variable explicative.*

L'objectif dans ce contexte est de caractériser l'aspect aléatoire du phénomène correspondant à la variable Y, comme dans le cadre de la ressource pédagogique "Statistique descriptive 1". Cependant, la différence est qu'ici, il s'agit de le faire en s'appuyant sur la connaissance d'une potentielle liaison avec un autre phénomène (celui associé à la variable X).

2.2 Liaison statistique

En préambule, il est ici nécessaire d'introduire certains objets et certaines notions mathématiques, permettant d'expliciter ce qu'est une liaison statistique. Pour cela, on rappelle ce à quoi correspond à la notion de distribution.

Définition 2.2.1 (Distribution ou Loi de probabilité) *Pour une variable aléatoire quantitative continue X, dont les valeurs possibles sont dans un intervalle $\mathcal{I}$, la distribution correspond à la connaissance de la (densité de) probabilité de chaque valeur possible de X.*

La distribution décrit de manière théorique l'aléatoire d'un phénomène étudié. Suivant les contextes, on peut utiliser "distribution" ou "loi de probabilité" pour faire référence à la même définition. Cette définition reste imprécise par rapport à ce qu'est réellement la définition formelle d'une distribution, mais cette définition a l'avantage d'être plus simple et plus compréhensible. Etant donné l'imprécision de cette définition, dans le cadre de ce cours on pourra accepter que "connaître la distribution d'une variable aléatoire" est équivalent à "connaître la densité de probabilité" ou "connaître la fonction de répartition".

Lorsqu'on dispose de deux variables aléatoires, il est possible de définir une distribution conjointe pour le couple des variables (les deux variables simultanément).

Définition 2.2.2 (Fonction de répartition jointe) *Pour deux variables aléatoires X et Y, la fonction de répartition jointe est :*

$$F_{XY}(x,y) = \mathbb{P}(X < x, Y < y),$$

ce qui s'interprète comme la probabilité que la variable X soit inférieure à une valeur donnée x et à la fois que la variable Y soit inférieure à une valeur donnée y.

Définition 2.2.3 (Densité jointe) *La densité jointe correspond à la double dérivée partielle de la fonction de répartition :*

$$\begin{aligned} f_{XY}(x,y) &= \frac{\partial}{\partial x}\left(\frac{\partial}{\partial y}F_{XY}(x,y)\right) \\ &= \frac{\partial^2}{\partial x \partial y}F_{XY}(x,y). \end{aligned}$$

Remarque 2.2.4 (Interprétation d'une densité jointe)**.** *Comme pour le cas avec une seule variable, pour lequel est aussi définit des notions de fonction de répartition et de fonction de densité de probabilité, dans le cas d'une densité de probabilité pour un couple de variables, la valeur de cette fonction s'interprète comme une intensité de probabilité locale. Autrement dit, si on a $f(x_1,y_1) > f(x_2,y_2)$, il sera plus vraisemblable d'observer des valeurs proches de (x_1,y_1) que proches de (x_2,y_2). La densité jointe indique avec quelle intensité il est vraisemblable d'observer à la fois la valeur x pour la variable X et à la fois la valeur y pour la variable Y.*

Remarque 2.2.5. *Le terme $\frac{\partial}{\partial x}$ correspond au faite d'effectuer un calcul de dérivée en ne considérant que x comme étant une variable. Autrement dit, même si dans l'expression à dériver il y a une autre variable (comme y), celle-ci sera considérée comme une constante du point de vue de la dérivation partielle par rapport à x. Des détails supplémentaires sont à retrouver dans les ressources pédagogiques de mathématiques, mais voici ci-dessous des exemples de résultats de la dérivation partielle donnant une illustration de ce dont il est fait mention au début de cette remarque :*

$$\begin{aligned} \frac{\partial}{\partial x}\left(x^2+3x+xy+2y\right) &= 2x+3+y \\ \frac{\partial}{\partial y}\left(x^2+3x+xy+2y\right) &= x+2 \\ \frac{\partial^2}{\partial x \partial y}\left(x^2+3x+xy+2y\right) &= 1 \end{aligned}$$

Une autre notion de densité importante est la densité conditionnelle, qui correspond à une intensité de probabilité d'une variable aléatoire, conditionnellement au fait qu'on fixe une valeur d'une autre variable aléatoire.

Définition 2.2.6 (Densité conditionnelle) *Pour une valeur donnée x_0, on définit la fonction de densité conditionnelle de Y sachant que $X = x_0$ de la manière suivante :*

$$f_{Y|X=x_0}(y) = \frac{f_{XY}(x_0,y)}{\int f(x_0,t)\,\mathrm{d}t}.$$

Cette fonction donne les intensités de probabilité pour chacune des valeurs y possibles, si on considère que la variable X est fixée à être égale à la valeur x_0. Pour se construire plus simplement une représentation mentale de ce que représente cette définition, la définition ci-dessous indique la version de la densité conditionnelle pour des variables quantitatives discrètes.

Définition 2.2.7 (Fonction de probabilité conditionnelle) *Pour une valeur donnée x_0, on définit la fonction de probabilité conditionnelle (ou fonction de masse conditionnelle) de Y, sachant qu'on a $X = x_0$, de la manière suivante :*

$$f_{Y|X=x_0}(y) = \mathbb{P}(Y = y|X = x_0) = \frac{\mathbb{P}(Y = y \cap X = x_0)}{\mathbb{P}(X = x_0)}.$$

Pour ce qui suit, les notions sont le plus souvent données dans le contexte de variables quantitatives continues (donc avec la notion de densité), mais le lecteur peut se construire les raisonnements et les formules nécessaires pour le cas de variables discrètes.

Exemple 2.2.8. *Pour se faire une idée concrète de ce à quoi correspond une densité conditionnelle, il convient d'utiliser un exemple simple et parlant. Prenons le cas de la relation qui existe entre le poids et la taille pour un adulte masculin. Pour cet exemple on va supposer qu'on connait parfaitement la population relative à ces caractérisques, autrement dit on connait les quantités théoriques (distributions et densités de probabilité). Supposons que les graphiques de la figure 2.1 correspondent aux densités de probabilité de la taille des individus (graphique de gauche) et de leurs poids (graphique de droite). De plus, les graphiques de la figure 2.2 représentent dans*

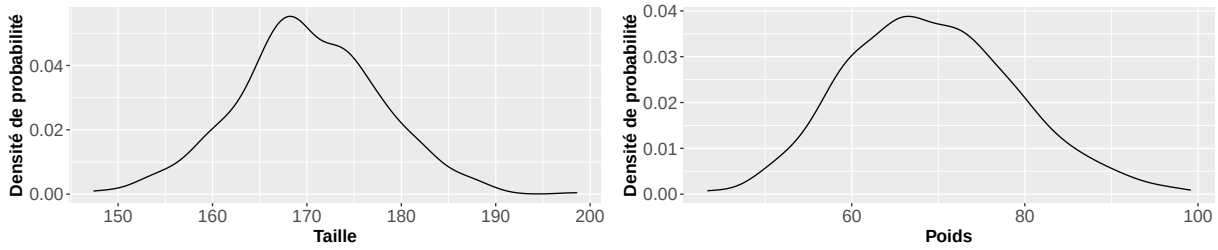


Figure 2.1 – *Les densités de probabilité des grandeurs taille (à gauche) et poids (à droite).*

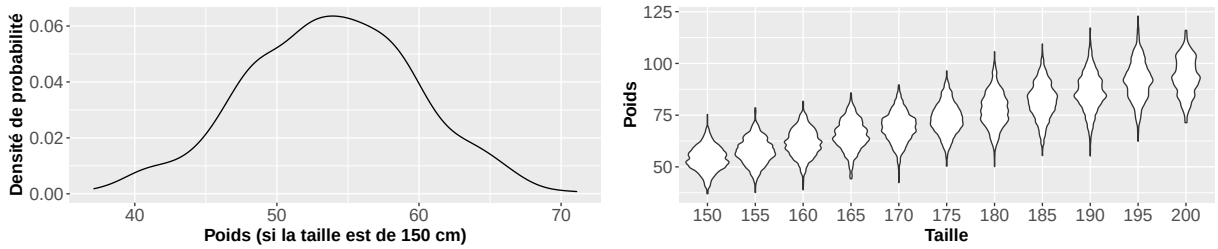


Figure 2.2 – *Densités conditionnelle du poids, conditionnellement à la taille. Le graphique de gauche est la densité de poids conditionnellement à des individus faisant 150cm. Le graphique de droite donne les densités de probabilité pour plusieurs niveaux de taille successifs allant de 150cm à 200cm.*

ce contexte ce qu'est une densité conditionnelle. Sur le graphique de gauche, il s'agit de la densité de probabilité du poids, conditionnellement à avoir une taille de 150cm. Pour l'interpréter, cela correspond à l'intensité de répartition théorique des poids possibles en supposant qu'on ne s'intéresse qu'à des individus ayant une taille de 150cm. On constate en comparant ce graphique avec la densité de probabilité du poids de toute la population (graphique de droite de la figure 2.1) que lorsqu'on fait 150cm, la variation de poids ne couvre que l'intervalle $[40, 70]$ *alors que pour l'ensemble des tailles possibles l'intervalle des valeurs possibles couvre l'intervalle* $[40, 100]$*. De plus, conditionnellement au faite de mesurer 150cm, le poids moyen est environ de 55kg alors que le poids moyen de toute la population des adultes masculins est plutôt de 70kg. Le graphique de droite montre plus globalement les densités conditionnelles du poids (conditionnellement à la taille) qu'on obtient pour différentes valeurs de tailles fixées. Dans cette exemple, on peut constater que l'ensemble des densités conditionnelles permet d'obtenir une connaissance quant à la liaison entre les deux variables.*

Pour compléter cet exemple, la figure 2.3 donne la densité jointe du couple de variables "taille-poids". Il est possible avec ce graphique de retrouver grossièrement ce à quoi ressemblent les densités conditionnelles en regardant ce graphique sur une ligne droite seulement (verticalement pour avoir un conditionnemment par rapport à la taille, et horizontalement pour avoir un conditionnement par rapport au poids).

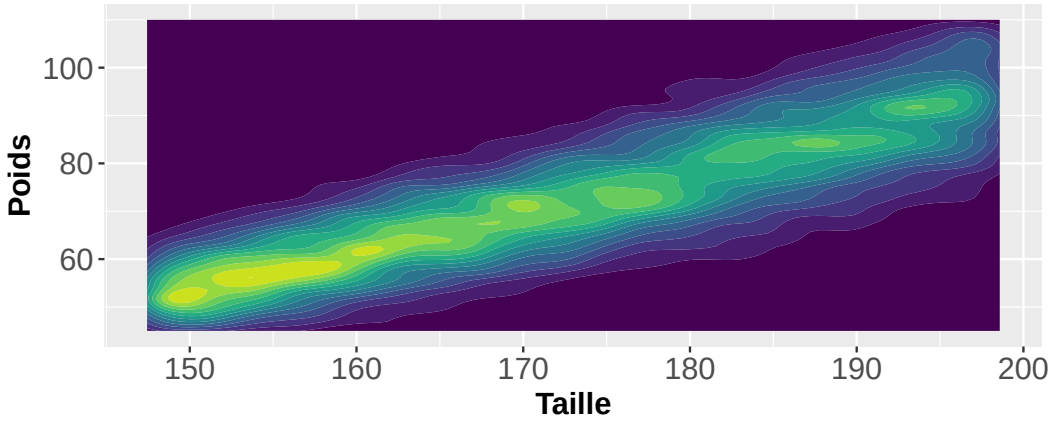


Figure 2.3 – *Densité jointe du poids et de la taille des individus. Le dégradé de couleur permet de représenter des échelles de densité de probabilité. Plus la couleur est sombre, plus la densité associée est faible, et inversément pour la couleur qui tend vers le clair.*

Comme on peut le constater avec l'exemple 2.2.8 et surtout avec le graphique de droite de la figure 2.2, la notion de densité conditonnelle a un lien avec l'idée qu'on peut avoir de ce qu'est une liaison statistique entre deux variables (quantitatives). Pour commencer à appréhender correctement cette notion, on définit ci-dessous avec la définition 2.2.9 ce qu'est l'indépendance entre deux variables aléatoires, autrement dit une absence de liaison statistique.

Définition 2.2.9 (Indépendance de deux variables aléatoires) *Deux variables aléatoires quantitatives continues X et Y sont dites indépendantes si elles vérifient :*

$$f_{XY}(x,y) = f_X(x)f_Y(y)$$

où f_X et f_Y sont les densités de probabilité respectives de X et Y.

Si les variables X et Y sont discrètes, alors l'équation d'indépendance est :

$$f_{XY}(x,y) = \mathbb{P}(X = x \cap Y = y) = \mathbb{P}(X = x)\mathbb{P}(Y = y).$$

Cette définition d'indépendance est la version en termes de densité, d'une définition de l'indépendance plus commune en termes de probabilités, voir ci-dessous.

Définition 2.2.10 (Indépendance de deux événements) *Deux événements A et B sont dites indépendantes si :*

$$\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B).$$

Pour donner une interprétation plus simple à l'équation de la définition 2.2.9 de l'indépendance entre deux variables aléatoires, on donne ci-dessous une propriété qui découle de cette définition.

Proposition 2.2.11 (Indépendance) *Si deux variables aléatoires continues X et Y sont indépendantes, alors l'équation suivante est vérifiée, pour toute valeur de x :*

$$f_{Y|X=x}(y) = f_Y(y)$$

Par analogie, si les variables sont discrètes, on a :

$$f_{Y|X=x}(y) = \mathbb{P}(Y = y|X = x) = \mathbb{P}(Y = y)$$

Cette proposition indique que même si on suppose connaître la valeur de la variable X, et qu'elle prend la valeur x, les probabilités (ou plutôt les densités) de chacune des valeurs possibles y ne changent pas. Avoir une information sur X, n'induit aucun changement sur l'aléatoire de la variable Y.

Exemple 2.2.12. *En reprenant le même contexte que l'exemple 2.2.8, on peut constater grace à la figure 2.2, qu'il n'y a pas indépendance entre les deux variables "taille" et poids". S'il n'y a pas d'indépendance, c'est donc qu'il y a une liaison statistique entre ces deux variables. Pour donner une exemple d'une paire de variables indépendantes, voici avec la figure 2.4 ce qu'on pourrait obtenir en termes de densité de probabilité jointe.*

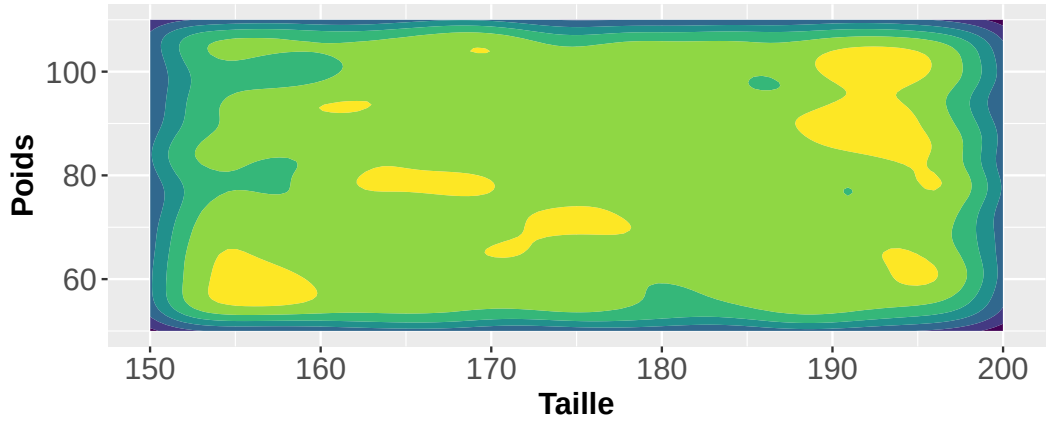


Figure 2.4 – *Densité jointe du couple "taille-poids" s'il y avait une indépendance entre ces deux grandeurs. Le code couleur est le même que celui de la figure 2.3 et on constate que quelque soit la taille, toutes les valeurs de poids sont possibles et elles sont de probabilités/densités similaires.*

Grace à ce qu'on peut définir comme étant une indépendance entre deux variables aléatoires, cela permet d'en déduire une définition de ce qu'est au contraire une dépendance, ou autrement dit une liaison statistique.

Définition 2.2.13 (Dépendance ou Liaison statistique) *Il y a une liaison statistique entre deux variables aléatoires X et Y dès lors que :*

$$f_{XY}(x, y) \neq f_X(x) f_Y(y).$$

Cette définition est très générale, ce qui permet de couvrir toutes les possibilités de dépendance statistique entre deux variables aléatoires. Cependant, en pratique on ne dispose pas des densités de probabilités, et encore moins de la densité jointe. Il est alors nécessaire d'appréhender ces quantités théoriques par des versions empiriques, mais dans la plupart des contextes réels ces versions empiriques ne sont pas assez précises pour vérifier correctement s'il y a égalité entre la densité jointe empirique $\hat{f}_{XY}$ et le produit des densités empiriques $\hat{f}_X \times \hat{f}_Y$. A noter que si on dispose d'une grande quantité de données, ces densités empiriques peuvent être très proches des leurs versions théoriques, cependant vérifier l'égalité de la définition 2.2.9 reste un problème complexe. Ainsi, il est plus commun de souhaiter vérifier une version plus simple de ce que pourrait être une liaison statistique.

On définit ci-dessous, de nouvelles notions permettant d'appréhender une liaison statistique sous des angles plus simples, mais qui au moins peuvent être abordables en pratique. Au lieu de caractériser une liaison à partir des densités, on peut s'intéresser à une caractéristique plus simple de la distribution théorique des données : l'espérance.

Définition 2.2.14 (Espérance conditionnelle) *L'espérance conditionnelle de Y sachant la valeur de X est la valeur de l'espérance de Y si on se restreint à la sous-population ayant la valeur x_0 pour la variable X.*

Dans le cas d'une variable aléatoire quantitative discrète, cette espérance a pour expression :

$$\mathbb{E}(Y|X = x_0) = \sum_{y \in \mathcal{Y}} y\, \mathbb{P}(Y = y|X = x_0)$$

où $\mathcal{Y}$ est l'ensemble des valeurs possibles de la variable Y.

S'il s'agit d'une variable aléatoire quantitative continue, l'espérance conditionnelle a pour expression :

$$\mathbb{E}(Y|X = x_0) = \int_{\mathcal{Y}} y\, f_{Y|X=x_0}(y)\, \mathrm{d}y$$

A l'aide de cette définition, il est possible de repenser la dépendance statistique, en la formulant ainsi : si la valeur de l'espérance conditionnelle n'est pas la même pour différentes valeurs x_1 et x_2 de X

$$\mathbb{E}(Y|X = x_1) \neq \mathbb{E}(Y|X = x_2)$$

alors il y a une liaison (en espérance) entre les variables X et Y. Autrement dit, si la fonction suivante

$$g(x) = \mathbb{E}(Y|X = x)$$

n'est pas une fonction constante, alors les deux variables ne sont pas indépendantes.

Avec ce point de vue, il est possible d'étudier à partir des données, une version empirique de cette espérance conditionnelle $g(x)$ et donc d'en déduire une potentielle liaison ou non entre les deux variables. Pour autant, définir un estimateur de cette fonction g n'est pas une chose simple et il faut utiliser des approches complexes, qui ne consistent pas simplement à calculer une moyenne pour estimer une espérance conditionnelle.

Définition 2.2.15 (Régression) *On appelle régression l'ensemble des méthodes statistiques ayant pour objectif de déterminer une estimation $\hat{g}$ de l'espérance conditionnelle d'une variable Y, sachant la valeur x d'une variable X.*

Cependant, à des fins pédagogiques il est préférable de ne pas commencer par introduire les outils nécessaires à cette étude (la régression) et on préfère ici se restreindre à un type particulier de fonction g (sous-cas de la régression), à savoir les fonctions linéaires. La section suivante détaille justement ce qu'il y a à savoir concernant la régression linéaire.

Remarque 2.2.16. *On parle ici de fonctions linéaires alors qu'il s'agit en réalité de fonctions affines. On s'autorise généralement cette confusion pour indiquer de manière intuitive qu'on considère les fonctions "qui sont des équations de droites/lignes" (ligne $\rightarrow$ linéaire). Il ne faut donc pas confondre cela avec l'emploi de "fonction linéaire" dans d'autres modules de mathématiques, dans lesquels "linéaire" fait référence à la propriété de linéairité ($f(\lambda x) = \lambda f(x)$).*

2.3 Liaison linéaire

Parmi les liaisons possibles en termes d'espérance conditionnelle entre des variable X et Y, la plus simple est la liaison linéaire.

Définition 2.3.1 (Liaison linéaire) *Il y a une liaison linéaire entre X et Y si il existe μ et β, deux paramètres tels que :*

$$g(x) = \mathbb{E}(Y|X = x) = \mu + \beta x.$$

L'expression $g(x) = \mu + \beta x$ correspond bien à une équation de droite, où μ correspond à l'ordonnée à l'origine, et β est le coefficient directeur de la droite. En pratique, pour étudier la potentielle liaison linéaire entre les deux variables, il est nécessaire de calculer une estimation des paramètres qu'on ne connaît pas, μ et β, ce qui est traité dans le chapitre 3.

Définition 2.3.2 (Sens de la liaison) *On dit qu'il y a une liaison positive s'il y a une liaison linéaire relative à une droite croissante. Autrement dit, si on a $\beta > 0$. Inversément, une liaison est dite négative si $\beta < 0$.*

Les graphiques de la figure 2.5 donnent des exemples de ce que peut être la densité jointe s'il y a une liaison positive ou négative entre deux variables.

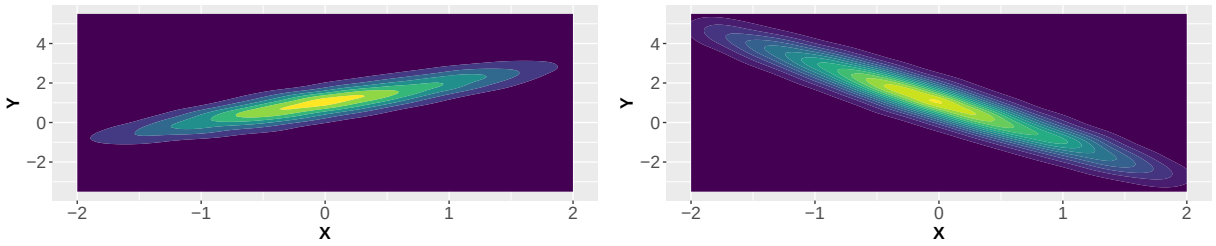


Figure 2.5 – *Exemples de densités jointes de couples de variables pour lesquelles il y a une liaison positive (graphique de gauche) ou une liaison négative (graphique de droite). Le code couleur est le même que celui de la figure 2.3.*

Définition 2.3.3 (Régression linéaire) *On appelle régression linéaire l'ensemble des méthodes statistiques ayant pour objectif de déterminer une estimation $\hat{\mu}$ et $\hat{\beta}$ de sorte à déterminer une liaison linéaire entre X et Y, en termes d'espérance conditionnelle.*

Une liaison linéaire est le type de liaison le plus communément étudié, du fait de sa simplicité, mais aussi pour le caractère interprétable d'une liaison linéaire. En effet, l'interprétation peut être la suivante : si il existe une liaison linéaire positive entre deux variables X et Y, cela signifie que si l'une augmente, l'autre augmente aussi (ou diminue s'il s'agit d'une liaison négative). De même, si l'une diminue, l'autre diminue aussi (ou augmente pour une liaison négative).

2.4 Liaison non-linéaire

Une liaison non-linéaire, en opposition à une liaison linéaire, correspond à une liaison qui ne peut pas se résumer en une équation de droite, ce qui induit généralement la nécessité de traitements statistiques plus complexes.

Définition 2.4.1 (Liaison non-linéaire) *Il y a une liaison non-linéaire entre X et Y si*

$$g(x) = \mathbb{E}(Y|X = x)$$

n'est pas une fonction qui peut s'écrire comme une équation de droite ($g(x) \neq \mu + \beta x$) et s'il ne s'agit pas d'une fonction constante ($g(x) = a$, où a est une constante).

En pratique dans cette ressource pédagogique, la notion de non-linéarité recouvre les cas où la fonction g peut être de forme exponentielle, logarithmique, polynomiale ou d'autres familles de fonctions usuelles, pour des cas anecdotiques.

Définition 2.4.2 (Régression non-linéaire) *On appelle régression non-linéaire l'ensemble des méthodes statistiques ayant pour objectif de déterminer une estimation de la fonction $g(x) = \mathbb{E}(Y|X = x)$ de sorte à déterminer une liaison non-linéaire entre X et Y, en termes d'espérance conditionnelle.*

Les graphiques de la figure 2.6 donnent des exemples de ce que peut être la densité jointe s'il y a une liaison non-linéaire entre deux variables.

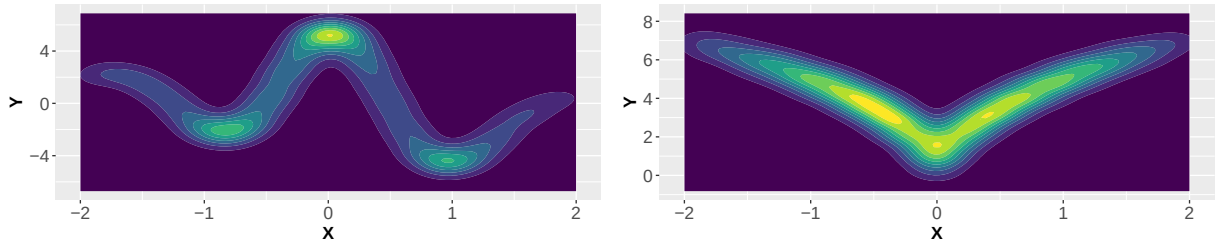


Figure 2.6 – *Exemples de densités jointes de couples de variables pour lesquelles il y a une liaison non-linéaire. Le code couleur est le même que celui de la figure 2.3.*

2.5 Liaison statistique et causalité

Lorsqu'on cherche une potentielle liaison entre deux variables aléatoires, c'est souvent parce qu'on cherche à déterminer un lien de cause à effet entre les deux phénomènes associés.

Définition 2.5.1 (Causalité) *Il y a une relation de causalité entre deux phénomènes A et B si, l'un est la conséquence de l'autre, à savoir si la réalisation d'un des deux phénomènes est conditionnée par l'état de l'autre phénomène.*

Exemple 2.5.2 (Pluie et parapluie)**.** *Un exemple simple de causalité dans la vie quotidienne est celui du lien entre les prévisions météos et le fait de prendre un parapluie avant de partir de chez soi. Dans le tableau suivant, on a les probabilités pour une personne prise au hasard de prendre son parapluie dans le cas où de la pluie est annoncée, et dans le cas où il n'y a pas de pluie annoncée.*

	Prendre un parapluie	*Ne pas prendre de parapluie*
Pluie annoncée	0.9	0.1
Pluie non-annoncée	0.01	0.99

Il paraît cohérent dans cet exemple simple, de dire qu'il y a un lien de cause à effet de "la pluie annoncée ou non" sur "le fait ou non de prendre un parapluie".

Exemple 2.5.3 (Débit d'une rivière : causalité vraisemblable)**.** *Autre exemple, lorsqu'on étudie les débits en amont et en aval d'une rivière, on imagine que ce qu'on observe sur le débit en aval est une conséquence de ce qu'on observe en amont, entre autres. Cependant, même si pour cet exemple, on peut avoir une intuition solide quant au lien de cause à effet entre ces deux grandeurs, dans des cas plus complexes, il n'est pas évident de statuer.*

Exemple 2.5.4 (Famille monoparentale et taux de criminalité : causalité à étudier)**.** *Pour donner de la nuance aux situations pour lesquelles on souhaite évaluer la présence d'une causalité, on peut prendre l'exemple d'une étude qui recense des indicateurs sociaux-économiques concernant une population. Si pour cette étude on établit une liaison entre le taux de criminalité avec le taux de famille monoparentale, on aurait tendance à y voir une lien de causalité. Autrement dit, on pourrait à tort faire le raccourci qu'être dans une famille monoparentale induit une hausse sur le taux de criminalité. Dans ce genre de cas, le but d'une analyse pertinente ne serait pas de se contenter de mettre en lumière cette liaison statistique, mais plutôt d'évaluer si derrière cette liaison statistique il y a effectivement un lien de cause à effet.*

Exemple 2.5.5 (Lunettes de soleil et crème solaire : causalité peu vraisemblable)**.** *Pour continuer, on peut donner un exemple pour lequel il est clair qu'il n'y a pas de causalité bien qu'on puisse détecter une liaison statistique. On considère les deux grandeurs suivantes : le nombre de lunettes de soleil vendues par jour et la quantité de crème solaire vendue par jour. On constate clairement en pratique une liaison statistique entre ces deux grandeurs, pour autant il n'y a pas de causalité. Autrement dit, on ne peut raisonnablement pas penser que c'est parce que beaucoup ou peu de lunettes de soleil se vendent que cela induit une modification sur la propension à l'achat de crème solaire. Dans ce cas-là, c'est plutôt qu'il y a un facteur extérieur qui influence simultanément les deux quantités en question. Plus précisément, il s'agit de la saison, et qu'on soit en hiver ou en été, cela détermine grandement les achats respectifs de lunettes de soleil et de crème solaire.*

Avec ces précédents exemples, on peut constater qu'il y a plusieurs types de causalité possibles. La figure 2.7 présente les différents schémas de causalité possibles. Les paragraphes ci-dessus donne de plus une alerte qu'il faut garder en tête : "la causalité est un concept différent de la liaison statistique". Non seulement il est plus complexe, mais surtout l'un n'implique pas forcément l'autre. Une présence de liaison statistique n'implique pas forcément un lien de causalité. Un lien de causalité n'implique pas forcément la présence d'une liaison statistique. Lorsqu'on interprète des résultats statistiques, il est donc recommandé de ne pas parler de causalité, mais plutôt de liaison ou d'association, statistique.

Pour autant, dans certains cas il est possible de pouvoir affirmer avec une bonne confiance qu'une liaison statistique est relative à une relation sous-jacente de causalité. Un premier cas où cela peut arriver, est lorsqu'on dispose des résultats d'une multitude d'études réalisées dans des populations potentiellement différentes, et avec des conditions expérimentales potentiellement différentes. Si l'ensemble de ces études détermine une liaison statistique entre deux phénomènes, cela contribue à renforcer notre croyance quant à une potentielle causalité.

Un deuxième cas est celui d'études pour lesquelles les données sont obtenues grace à un *plan d'expérience* dédié à brouiller l'influence de variables extérieures qui ne sont pas mesurées. Le terme de plan d'expérience correspond à la stratégie de collecte de données, et il y a des contextes pour lesquels les expérimentateurs peuvent fixer certaines conditions. Par exemple, pour une étude cherchant à établier un lien entre l'activité physique et la probabilité de déclarer une pathologie cardiorespiratoire, il est possible d'imposer aux personnes participant à l'expérience un niveau d'activité physique prédéfini. Cela permet d'éviter que les deux grandeurs puissent être déterminées par une autre variable extérieure. Par exemple, le faite d'être fumeur pourrait avoir un lien avec ces deux grandeurs : fumer tend à diminuer le niveau d'activité physique et augmente le risque de contracter une pathologie cardiorespiratoire. En imposant un niveau d'activité physique, les expérimentateurs permettent de découpler l'activité physique du statut de fumeur, ce qui permet d'augmenter la confiance qu'on aurait en considérant qu'une liaison statistique pourrait être en réalité une causalité dans ce cas-là.

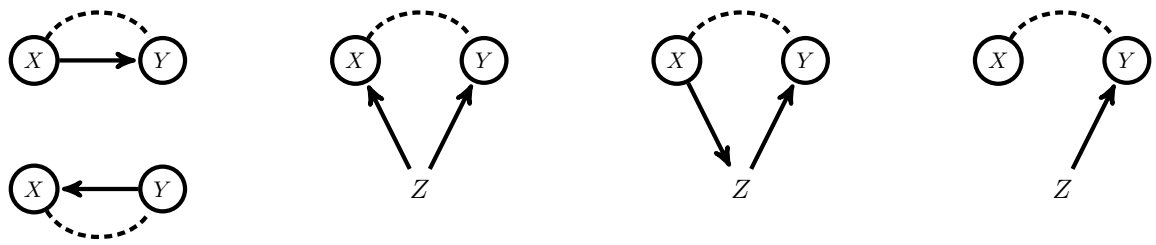


Figure 2.7 – *Les différentes structures de causalité (flèche en trait plein) et de liaison statistique (trait en pointillé). Autrement dit, pour chacun des cas présentés, on détecte une liaison statistique entre les variables X et Y. Les diagrammes de gauche illustrent un cas où il y a un lien de causalité, X est la cause de Y ou inversément. Pour le deuxième diagramme, il y a une variable (cachée) qui n'est pas observée et qui est la cause des deux variables X et Y. Cette dépendance commune à Z induit une liaison statistique artificielle entre X et Y. Pour le troisième diagramme, le schéma de causalité va de X à Y en passant par Z ce qui induit une liaison statistique artificielle entre X et Y. Pour celui de droite, la cause de Y est Z mais bien qu'il n'y ait pas de lien causal entre X et Y, il se peut qu'on puisse tout de même déterminer en pratique une liaison statistique (à tort) entre X et Y.*

Un troisième cas de détection de causalité est celui d'une étude statistique déterminant une association entre deux grandeurs, et qui est a posteriori confirmée avec une approche non-statistique. Par exemple, si on constate statistiquement que l'activité physique induit une diminution du risque de pathologie cardiorespiratoire, cela peut donner une piste à explorer d'un point de vue bio-médical. Et si, au terme de cette piste, un modèle biologique des mécanismes en jeu dans l'activité physique et des pathogologies cardiorespiratoires, tend à montrer que l'activité physique consolide les organes et les artères en lien avec le système cardiorespiratoire, alors on aura décelé un lien de causalité. Pour le formuler autrement, détecter une liaison statistique est un des indices pour guider de nouveles études, et est donc seulement un des éléments contribuant à la détermination d'une causalité.

CHAPITRE 3

Ajustement linéaire

Dans ce chapitre, comme pour le suivant, il est question de l'ajustement d'un modèle. En particulier, il s'agit ici du cas de l'ajustement de modèle linéaire et pour le chapitre 4 d'ajustement de modèles non-linéaires. Avant d'aller plus loin, on définit ce qu'est un modèle et ce qu'est l'ajustement d'un modèle.

Définition 3.0.1 (Modèle) *On parle de modèle (ou modèle probabiliste) lorsqu'on utilise une décomposition suivante pour analyser des données :*

$$\textit{données} = \textit{modèle} + \textit{variations}$$

qui correspond à décrire les données observées (qui sont considérées comme aléatoires) au travers d'une partie non-aléatoire et calculable à partir d'une équation mathématique (modèle), et d'une partie imprévisible et qui correspond à une quantité spécifique à chaque individu (variations).

Définition 3.0.2 (Ajuster) *On ajuste un modèle à des données lorsqu'on calibre le ou les paramètres du modèle, de sorte à ce que le modèle donne une description fidèle des données.*

En pratique, pour déterminer s'il y a une liaison statistique entre deux grandeurs, la procédure consiste 1) à déterminer un modèle adapté pour les données, 2) à ajuster un modèle et 3) à déterminer si le modèle ajusté indique qu'il y a une liaison ou non.

Exemple 3.0.3 (Prise de masse)**.** *On s'intéresse à une population de personnes âgées, celles souffrant de troubles du métabolisme, ce qui a pour conséquence que ces personnes ont du mal à prendre de la masse (masse graisseuse ou masse musculaire). Ces troubles peuvent engendrés une difficulté à récupérer après une maladie ou une opération. On considère ici un programme expérimental ayant pour objectif d'améliorer la prise de masse de ces personnes. Dans le cadre de l'analyse de ce protocole expérimental, on stipule que la prise de masse correspond à un pourcentage de la masse initiale de l'individu, mais on ne connait pas ce pourcentage, qu'on notera θ. D'après l'hypothèse qu'on cherche à étudier, on peut écrire la relation suivante :*

$$p = (1 + \theta) \times p_0 \tag{3.1}$$

où p est le poids à la fin de l'expérience, et p_0 est le poids initial. L'écriture de l'équation (3.1) revient en réalité à définir un modèle théorique de l'effet du traitement sur la masse des individus. Mais en l'état, on ne sait pas si ce modèle est valide, ni même quelle serait la valeur potentielle du paramètre θ. Pour cela, il faut ajuster le modèle, à savoir faire en sorte de trouver une valeur de θ qui permettent de décrire effectivement les prises de masses personnes individus de l'étude.

Cependant, en consultant les données, on se rend compte qu'il y a un des individus qui faisait 63kg et qui est passé à 64kg, soit une prise de masse de 1.59%, alors qu'un autre individu est passé de 50kg à 52.5kg, ce qui correspond à une prise de masse de 5%. Cela semble aller en contradiction avec l'écriture du modèle qu'on a donné avec l'équation (3.1), puisqu'il ne semble pas y avoir en réalité un seul pourcentage de prise de masse θ mais potentiellement autant de pourcentages que d'individus. Pire encore, au terme de l'expérience un individu a perdu un pourcentage de 1.02% de son poids. Il y a des variations pour chaque individus et elle peuvent aller jusqu'à des valeurs négatives. Prise isolément, chaque donnée ne permet pas de manière évidente de statuer sur l'efficacité ou non de ce traitement expérimental.

Pour modéliser avec plus de flexibilité ce paramètre de pourcentage de prise de masse, on va s'autoriser à supposer qu'il y a des variations dans l'expérience, qui peuvent être dues à des spécificités individuelles ou à des

facteurs extérieurs qu'on ne peut pas contrôler dans ce contexte expérimental. On suppose donc que le poids à la fin de l'expérience correspond à un pourcentage de prise de masse, plus ou moins une variation aléatoire relative à chaque individu :

$$p = (1 + \theta) \times p_0 + \varepsilon$$

et le terme ε correspond à cette variation individuelle du poids.

En utilisant des méthodes d'estimation, on peut calculer à partir des données de l'expérience une estimation $\hat{\theta}$, qui correspond à un pourcentage global de prise de masse. Après calcul, on obtient $\hat{\theta} = 2.79\%$, ce qui semble nous indiquer que le groupe de personnes testées ont globalement pris de la masse, bien qu'il y ait des variations de prises de masse, et même s'il y a quelques individus qui ont perdu du poids. Au final, on a obtenu le modèle de prise de masse suivant pour étudier cette population :

$$p = 1.0279 \times p_0 + \varepsilon$$

et pour cela, on a ajusté le modèle aux données qu'on avait à disposition.

L'objectif de ce chapitre est de donner les concepts relatifs à un premier modèle standard (la régression linéaire), ainsi que de fournir les outils à utiliser pour réaliser par soi-même une analyse avec ce modèle. Cela permet de pouvoir répondre à des problématiques concrètes comme par exemple 1) de déterminer s'il y a une liaison linéaire entre deux grandeurs comme la taille et le poids d'une personne, 2) d'indiquer s'il s'agit d'une liaison positive ou négative, 3) d'être capable de prédire ce que devrait être le poids d'un individu d'une taille donnée, et 4) de détecter des données atypiques.

Dans la suite de ce chapitre, la section 3.1 introduit le modèle de régression linéaire, qui est central dans ce chapitre. La section 3.2 explique des fondements mathématiques de l'ajustement de ce modèle. Par la suite, le moyen d'évaluer la qualité de l'ajustement d'un modèle est donné en section 3.3. Pour finir ce chapitre, la section 3.4 contient les diapos de cours ainsi que les feuilles de TD.

Table des matières de ce chapitre

3.1 Régression linéaire simple

Pour un ensemble de n unités statistiques, on dispose de deux séries de données quantitatives continues. On les note $x = (x_1, \ldots, x_n)$ et $y = (y_1, \ldots y_n)$. L'objectif est de déterminer s'il y a une liaison statistique entre les variables X et Y, qu'on puisse déceler entre les deux séries de données.

Pour cela, on introduit le modèle de régression linéaire simple, qui donne une modélisation mathématique de la potentielle liaison entre les deux variables.

Définition 3.1.1 (Régression linéaire simple) *Le modèle de régression linéaire simple correspond à l'équation suivante, qui exprime les valeurs y_i comme étant calculables à partir des données x_i :*

$$y_i = \mu + \beta x_i + \varepsilon_i \tag{3.2}$$

où μ et β sont les paramètres d'une équation de droite. La quantité ε_i symbolise l'erreur du modèle, il s'agit de la variation théorique des données autour de la modélisation.

Remarque 3.1.2 (Simple). *L'adjectif de "simple" pour le modèle de régression linéaire simple, indique que pour expliquer et prédire la variable Y, on ne va utiliser qu'une seule variable explicative X. Dans des problématiques plus complexes, on peut vouloir utiliser simultanément plusieurs variables explicatives $X_1, X_2, \ldots, X_p$ pour expliquer et prédire la variable Y. Pour faire cela, il sera nécessaire d'employer une version différente du modèle de régression, qui ne sera pas la "régression linéaire simple".*

Ce modèle indique que chaque donnée y_i est liée par un calcul d'équation de droite à la donnée x_i. Cependant, pour que cette modélisation soit assez flexible pour bien décrire les paires de données (x_i, y_i), on s'autorise à ce qu'il y ait un écart entre le calcul de l'équation de droite, et la véritable valeur de y_i. Cela s'explique par le fait qu'il serait trop restrictif de supposer que les points de coordonnées (x_i, y_i) relatifs aux données, se répartissent exactement sur une ligne droite. Même pour une phénomène physique pour lequel on s'attend à trouver une liaison

linéaire parfaite entre deux grandeurs, étant donnée des variabilités non-maitrisables des capteurs de mesure (ce qui est inhérent à toute étude empirique), il sera normal d'obtenir des données dont les points ne se répartissent pas exactement sur une ligne droite.

Exemple 3.1.3 (Distance de freinage)**.** *Des données collectées dans les années 1920 concernent les distances de freinage des voitures de l'époque, en fonction de la vitesse de la voiture. Si on s'intéresse à étudier le lien entre ces deux grandeurs, la régression linéaire simple est une approche adaptée. Lorsqu'on trace le nuage de points relatif à ces données, on obtient le graphique de la figure 3.1 et on se rend compte qu'il n'est pas possible de trouver un lien de proportionnalité parfait (à savoir faire passer une droite par l'ensemble des points), bien que le phénomène physique en question soit plutôt simple.*

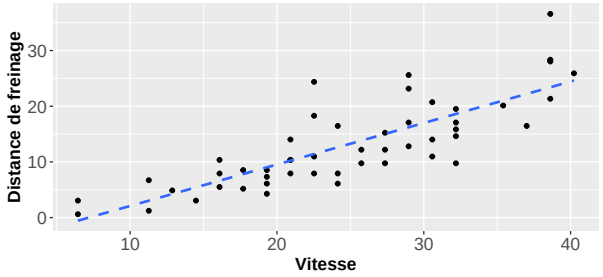


Figure 3.1 – *Nuage de points des données de distance de freinage de l'exemple 3.1.3. La droite en pointillé correspond à la meilleure droite possible pour passer "au travers de l'ensemble des points", et qui est obtenue par ajustement du modèle de régression linéaire simple.*

Pour ce modèle, qui fait intervenir une équation de droite ($\mu + \beta x_i$), le paramètre μ s'appelle "l'ordonnée à l'origine" (ou *intercept* en anglais) et le paramètre β s'appelle le "coefficient de pente" (ou *slope coefficient* en anglais). Chacun de ces deux paramètres a une interprétation possible :

- Le paramètre μ correspond à la valeur qu'on peut attendre de la variable Y si on a un individu i pour lequel $x_i = 0$. Pour s'en convaincre, il suffit d'utiliser l'équation $x_i = 0$ dans l'équation (3.2) du modèle de régression linéaire simple et on obtient :

$$y_i = \mu + \beta \times x_i + \varepsilon_i \quad \Longrightarrow \quad y_i = \mu + \beta \times 0 + \varepsilon_i \quad \Longrightarrow \quad y_i = \mu + \varepsilon_i,$$

  ce qui s'interprète comme : "d'après le modèle, la valeur de y_i devrait être environ égale à μ, à une erreur près qui correspond au terme ε_i". Généralement, cette interprétation n'est pas celle qu'on cherche à obtenir avec l'utilisation de ce modèle, et donc on ne s'intéresse généralement pas à la valeur du paramètre μ. Sa valeur sera principalement utile pour certains calculs.
- Le paramètre β est le coefficient directeur de la droite décrite par l'équation (3.2) du modèle. Une première interprétation calculatoire de ce coefficient β consiste à déterminer que si on augmente la valeur de x_i d'une valeur de h (on obtient alors une nouvelle donnée $x_{n+1} = x_i + h$), alors on devrait s'attendre à augmenter la valeur de la donnée y_{n+1} correspondante d'une valeur de $\beta \times h$.
  Une interprétation différente de ce coefficient concerne le lien qu'il y a entre le signe du coefficient et le sens de la liaison linéaire, voir la proposition 3.1.4.

Proposition 3.1.4 (Signe de β et sens de la liaison) *Le signe du coefficient β est en lien avec le sens de la potentielle liaison linéaire entre les deux variables :*

- *Si $\beta > 0$, alors la liaison linéaire est positive : quand une des deux variables augmente, l'autre augmente aussi d'une intensité correspondant à la valeur de β.*
- *Si $\beta = 0$, il n'y a pas de liaison linéaire : quand une des deux variables augmente, on ne constate pas de changement dans les valeurs de l'autre variable.*
- *Si $\beta < 0$, alors la liaison linéaire est négative : quand une des deux variables augmente, l'autre diminue d'une intensité correspondant à la valeur absolue $|\beta|$.*

La figure 3.2 donne une illustration de chacun des trois cas possibles de l'énoncé de la proposition 3.1.4.

En pratique pour utiliser cette approche, il faut être capable de déterminer les valeurs de μ et β. Comme ce sont des paramètres (qui sont donc inconnus), il est nécessaire d'utiliser des estimations de ces paramètres : $\hat{\mu}$ et $\hat{\beta}$. Pour obtenir les valeurs de ces estimations, il faut utiliser les expressions suivantes :

$$\hat{\beta} = \frac{\text{cov}(x,y)}{s_x^2} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2} \tag{3.3}$$

$$\hat{\mu} = \overline{y} - \hat{\beta}\overline{x} \tag{3.4}$$

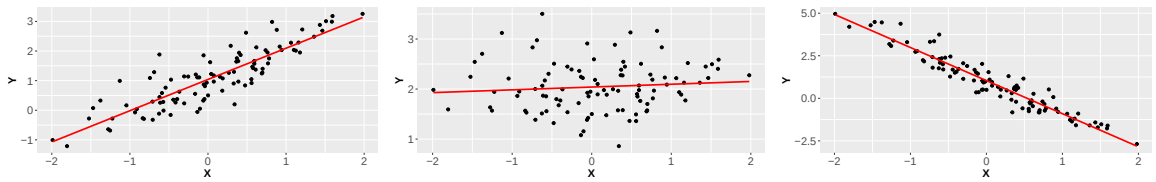


Figure 3.2 – *Des exemples de nuages de points (avec la droite du modèle de régression linéaire simple) pour des coefficients de pente β ayant des signes différents. Pour le graphique de gauche, la droite est croissante ce qui correspond à un coefficient de pente $\beta > 0$. Pour le graphique du milieu, la droite est presque horizontale donc $\beta \approx 0$. Pour le graphique de droite, la droite est décroissante donc $\beta < 0$.*

Proposition 3.1.5 (Simplification du calcul de $\hat{\beta}$) *De la même manière qu'il y a une simplification du calcul de la variance empirique (voir ressource pédagogique "Statistique descriptive 1"), il y a une simplification du calcul de l'estimateur $\hat{\beta}$ ayant pour formule (3.4) :*

$$\hat{\beta} = \frac{\overline{xy} - \overline{x}\,\overline{y}}{\overline{x^2} - \overline{x}^2}$$

où $\overline{xy} = \frac{1}{n}\sum_{i=1}^{n} x_i y_i$ et $\overline{x^2} = \frac{1}{n}\sum_{i=1}^{n} x_i^2$.

Proposition 3.1.6 (Lien entre $\hat{\beta}$ et le coefficient de corrélation) *Pour rappel, on a que $r(x,y) = \frac{\mathrm{cov}(x,y)}{s_x s_y}$ et donc on a :*

$$\hat{\beta} = r(x,y) \times \frac{s_y}{s_x}.$$

Autrement dit, à un coefficient près, l'estimation du coefficient de pente $\hat{\beta}$ correspond à la corrélation linéaire empirique entre les deux séries de données x et y. De plus, comme les termes s_x et s_y sont des termes nécessairement strictement positifs, l'estimation $\hat{\beta}$ est de même signe que $r(x,y)$.

Les interprétations concernant le signe du coefficent de pente estimé $\hat{\beta}$ sont donc les mêmes que celles du signe de la corrélation linéaire empirique $r(x,y)$.

Une fois que sont estimés les paramètres μ et β, on dispose des objets statistiques définit ci-dessous.

Définition 3.1.7 (Prédiction) *Pour une donnée x_i, la prédiction qu'on peut faire à l'aide du modèle de régression linéaire et des estimations des paramètres est :*

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

Qu'on connaisse ou non la valeur y_i, la prédiction $\hat{y}_i$ peut être utilisée pour de nombreux calculs autour de l'utilisation de ce modèle de régression. La figure 3.3 montre à quoi correspond graphiquement le calcul de cette prédiction. A savoir, calculer une prédiction pour une données x_i revient à déterminer le point de la droite dont l'abscisse est x_i. De plus, pour étudier un modèle comme la régression linéaire simple, une quantité importante est

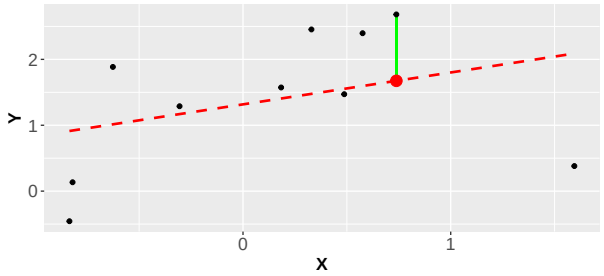


Figure 3.3 – *Représentation de ce qu'est une prédiction (point rouge) et une erreur de prédiction (trait vert) pour le modèle de régression linéaire simple. La ligne droite rouge en pointillé correspond à la droite qu'on obtient en ajustant le modèle de régression linéaire.*

le résidu, qui peut se comprendre comme la version empirique du terme d'erreur théorique ε_i du modèle.

Définition 3.1.8 (Résidu) *Le résidu e_i correspond à l'erreur de prédiction, à savoir à l'écart entre la prédiction et la vraie donnée :*

$$e_i = y_i - \hat{y}_i.$$

Proposition 3.1.9 (Moyenne des résidus) *Etant données les formules des estimateurs $\hat{\mu}$ et $\hat{\beta}$, on a :*

$$\overline{e} = \frac{1}{n}\sum_{i=1}^{n} e_i = 0.$$

Autrement dit, en moyenne on ne fait pas d'erreur de prédiction sur l'ensemble des données.

Définition 3.1.10 (Modèle de régression linéaire ajusté) *On appelle le modèle de régression linéaire ajusté, une expression similaire à celle de la formule (3.2) du modèle de régression linéaire, mais avec les versions empiriques des termes théoriques :*

$$y_i = \hat{\mu} + \hat{\beta} x_i + e_i.$$

Définition 3.1.11 (Droite de régression) *La droite de régression est la droite dont l'équation correspond à celle du modèle ajusté :*

$$y = \hat{\mu} + \hat{\beta} x.$$

On retrouve des exemples de droite de régression dans les graphiques des figures 3.2, 3.3 et 3.4.

Exemple 3.1.12. *Pour l'exemple de la régression de la taille des pères et des fils, introduit en section 1.2, on calcule ci-dessous, l'ensemble des quantités intervenant dans l'application du modèle de régression linéaire. Pour cette problématique, la donnée x_i est la taille du $i^{ème}$ père et la donnée y_i est la différence de taille avec son fils. On obtient les valeurs suivantes :*

- *$\overline{x} = 175.6872$ et $\overline{y} = 0.1540387$ (en moyenne il n'y a pas de différence de taille entre le groupe des pères et le groupe des fils),*
- *$\overline{xy} = 8.256522$ et $\overline{x^2} = 30900.03$,*
- *ce qui permet d'obtenir les estimations $\hat{\beta} = -0.5522521$, et $\hat{\mu} = 97.17764$.*

En arrondissant au centième pour faire simple, le modèle ajusté s'écrit ici comme :

$$y_i = 97.18 - 0.55 x_i + e_i$$

et la droite de régression est donnée avec la ligne rouge du graphique de la figure 3.4. Les types de résultats qu'on

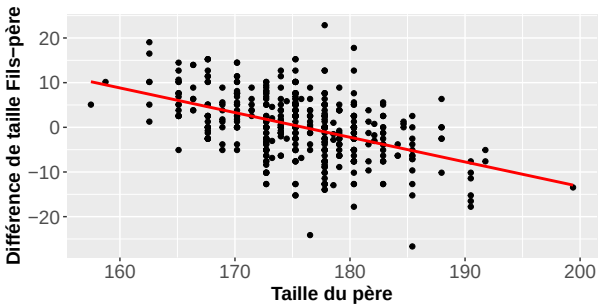


Figure 3.4 – *Nuage de points et droite de régression (en rouge) pour les données de taille père-fils.*

peut fournir suite à l'ajustement de ce modèle à ces données sont :

- *L'estimation du coefficient de pente $\hat{\beta}$ est négative ce qui tend à montrer que la taille du père et la différence avec la taille du fils sont associées négativement.*
- *Pour une taille de père donnée, par exemple de $x = 170cm$ (en dessous de la moyenne), alors on a une prédiction de $\hat{y} = 3.29cm$, ce qui implique que la taille prédite pour son fils est $173.29cm$. Par contre, pour une taille de père de $x = 190cm$ (au-dessus de la moyenne), on a une prédiction $\hat{y} = -7.75cm$ et donc une taille prédite pour le fils de $182.25cm$.*

Mise en pratique des notions de la section 3.1

Exercice 3.1.1 (Application)**.** Appliquez le modèle de régression linéaire simple sur les données du tableau ci-dessous :

i	x_i	y_i
1	-0.63	0.05
2	0.18	1.45
3	-0.84	-0.80
4	1.60	3.75
5	0.33	1.88
6	-0.82	-0.65
7	0.49	1.97
8	0.74	2.67
9	0.58	2.32
10	-0.31	0.51

Pour cela, il vous faut calculer les estimations $\hat{\beta}$ et $\hat{\mu}$. De plus calculez les prédictions du modèle pour ces 10 valeurs de x_i. Pour finir, calculez les résidus et vérifier que la moyenne des résidus est égale à 0.

Exercice 3.1.2 (Moyenne des résidus)**.** En utilisant la formule de l'estimateur (3.3) et la formule de la prédiction, définition 3.1.7, faites la démonstration de la proposition 3.1.9.

Exercice 3.1.3 (Démonstration des formules simplifiées)**.** Cet exercice a pour objectif de démontrer la simplification de la proposition 3.1.5. Pour cela, on commence par montrer la simplification de la variance empirique (vue dans la ressource pédagogique "Statistique descriptive 1").

1. Expliquer pourquoi $\frac{1}{n}\sum_{i=1}^{n}(x_i \times \overline{x}) = (\overline{x})^2$.
2. Expliquer pourquoi on peut distribuer l'opérateur de somme comme suit :

$$\sum_{i=1}^{n}(ax_i + b) = a\left(\sum_{i=1}^{n} x_i\right) + b$$

3. Développer le carré dans la formule suivante :

$$\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2$$

   puis distribuer l'opérateur de somme pour chacun des termes.
4. En déduire que $s_n^2 = \overline{x^2} - (\overline{x})^2$.

Pour finir la démonstration, il faut procéder de la même manière pour le dénominateur :

5. Développer le double produit $(y_i - \overline{y})(x_i - \overline{x})$.
6. Distribuer l'opérateur de somme.
7. En déduire que $\frac{1}{n}\sum_{i=1}^{n}(y_i - \overline{y})(x_i - \overline{x}) = \overline{xy} - \overline{x}\,\overline{y}$.

3.2 Méthode des moindres carrés

Dans cette section, il est question de la méthode qui sert à trouver les formules des estimateurs (3.3) et (3.4). Cela est aussi l'occasion d'introduire des concepts importants pour les approches statistiques par modélisation.

3.2.1 Intuition et principe de la méthode des moindres carrés

La question de déterminer les estimateurs du modèle de régression linéaire simple peut se réexprimer de la manière suivante : "trouver les valeurs $\hat{\mu}$ et $\hat{\beta}$ qui permettent d'obtenir les meilleurs prédictions". D'un point de vue géométrique, cela peut se reformuler ainsi : "trouver la meilleure droite de régression qui passe au travers du nuage de points". Pour cela, imaginons avoir deux paires d'estimations candidates pour être les meilleures estimations :

- la première paire est $\hat{\mu} = 1$ et $\hat{\beta} = 2$, et
- la seconde paire est $\hat{\mu} = 1.5$ et $\hat{\beta} = 1.5$.

La figure 3.5 donne une représentation graphique pour chacun de ces deux cas, en termes de nuage de points et de droite de régression. A l'œil nu il est n'est pas évident de statuer laquelle est la meilleure. Pour appréhender ce qui serait les meilleures estimations, on peut s'intéresser aux résultats qu'elles donnent en terme de prédictions. Plus précisément, est-ce que les prédictions $\hat{y}_i$ qu'on obtient sont proches des valeurs mesurées y_i ? De plus, est-ce celles

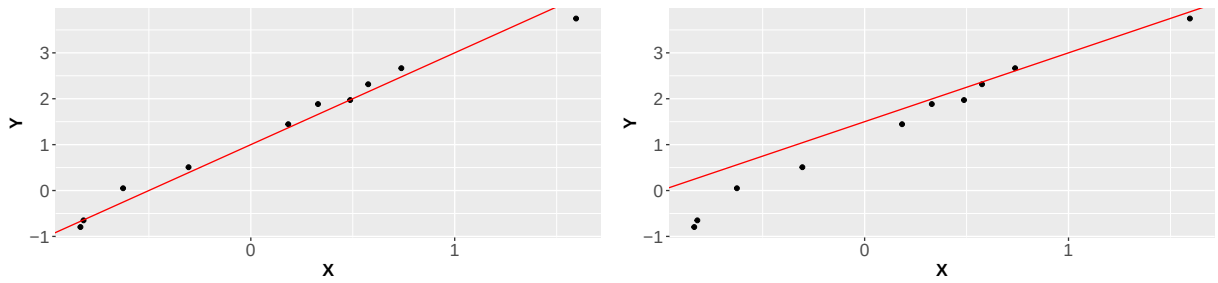


Figure 3.5 – *Différentes droites candidates pour être la meilleure droite de régression.*

obtenues avec la première paire ($\hat{\mu} = 1, \hat{\beta} = 2$) qui donnent les meilleures prédictions ou celles avec la seconde paire ($\hat{\mu} = 1.5, \hat{\beta} = 1.5$) ? Pour évaluer cela, commençons par calculer les prédictions pour une donnée : celle de coordonnées $(-0.82, -0.65)$. Dans ce cas, on a $y_i = -0.65$ et

$$\text{avec } \hat{\mu} = 1, \hat{\beta} = 2 \text{ la prédiction est } \hat{y}_i = 1 + 2 \times (-0.82) = -0.64 \text{ , d'où un écart de } 0.01,$$
$$\text{avec } \hat{\mu} = 1.5, \hat{\beta} = 1.5 \text{ la prédiction est } \hat{y}_i = 1.5 + 1.5 \times (-0.82) = 0.27 \text{ , d'où un écart de } 0.91.$$

Pour cette donnée, la première paire d'estimations donne une meilleure prédiction. Cependant, on constate l'inverse si l'on considère la donnée suivante : $(1.6, 3.74)$. Pour celle-ci on a $y_i = 3.74$ et

$$\text{avec } \hat{\mu} = 1, \hat{\beta} = 2 \text{ la prédiction est } \hat{y}_i = 1 + 2 \times (1.6) = 4.2 \text{ , d'où un écart de } 0.46,$$
$$\text{avec } \hat{\mu} = 1.5, \hat{\beta} = 1.5 \text{ la prédiction est } \hat{y}_i = 1.5 + 1.5 \times (-0.82) = 3.9 \text{ , d'où un écart de } 0.3.$$

Le raisonnement à avoir c'est qu'il faut prendre conjointement toutes les données pour choisir entre les deux paires d'estimations. Cela revient graphiquement à évaluer quelle droite de régression donne des erreurs (en rouge) les plus réduites, voir la figure 3.6. Pour calculer cela, on calcule la moyenne des écarts entre les données y_i et les

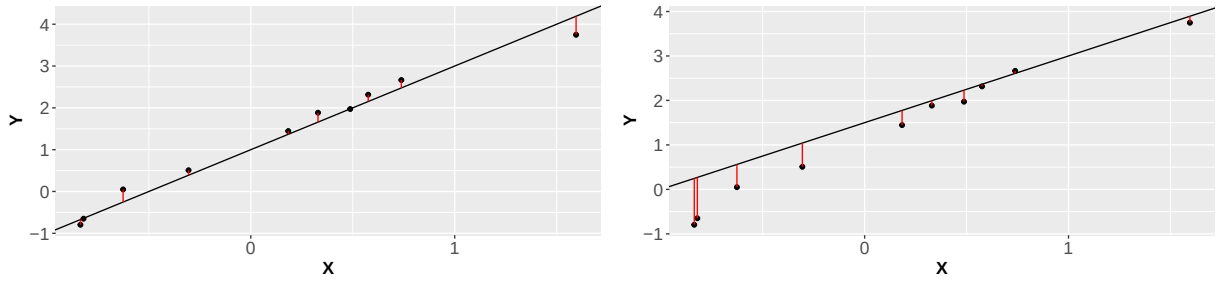


Figure 3.6 – *Ecarts entre les données y_i et les prédictions $\hat{y}_i$.*

prédiction $\hat{y}_i$, ce qui correspond au "critère des moindres carrés".

Définition 3.2.1 (Critère des moindres carrés) *Le critère des moindres carrés est :*

$$C(\theta) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2.$$

où θ est l'ensemble des paramètres du modèle en question.

Dans le cas du modèle de régression linéaire simple, on a $\theta = (\mu, \beta)$, et donc :

$$C(\mu, \beta) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - (\mu + \beta x_i))^2 = \frac{1}{n}\sum_{i=1}^{n} e_i^2 \tag{3.5}$$

Pour trouver les meilleurs estimations, il faut donc chercher à minimiser cette fonction, puisque cela correspond au fait de minimiser l'écart (au carré) entre les prédictions et les données, et autrement dit à minimiser les résidus (au carré).

Remarque 3.2.2 (Moindres carrés)**.** *On parle de "moindres carrés" puisqu'on cherche à minimiser les carrés des erreurs de prédiction (résidus).*

3.2.2 Minimiser une fonction à une variable

Chercher les estimations du modèle de régression linéaire se ramène donc à une problème de minimisation d'une fonction. On cherche donc un couple de valeurs (μ_0, β_0), tel que

$$\forall(\mu, \beta) \in \mathbb{R}^2, C(\mu_0, \beta_0) \leq C(\mu, \beta).$$

Il existe des procédures mathématiques pour déterminer le minimum d'une fonction, et pour commencer avec un cas illustratif simple, prenons l'exemple d'une fonction à une seule variable $f(x)$ qu'on cherche à minimiser. Si la fonction est $f(x) = x^2$, on aura que $x_0 = 0$ est un minimum de cette fonction. Pour s'en convaincre, on peut soit tracer le graphique de la fonction, soit montrer que $f(x) \geq 0, \forall x \in \mathbb{R}$ et que comme $f(0) = 0$, alors $x_0 = 0$ est bien un minimum de la fonction f. Tout comme pour cet exemple, on ne traite ici et dans la suite que des fonctions continues, dérivables et dont la dérivée est continue (autrement dit f est deux fois dérivable).

Dans le pararaphe précédent, on a déterminé avec l'intuition et à la main, le minimum d'une fonction, mais pour le trouver avec une méthode plus rigoureuse, on commence par déterminer les points critiques de la fonction.

Définition 3.2.3 (Point critique) *Pour une fonction f à une variable, un point critique est un abscisse x_0 qui vérifie :*

$$f'(x_0) = 0,$$

à savoir c'est un annulateur de la première dérivée de f.

La figure 3.7 donne une représentation graphique des différents types de points critiques possibles. Chercher un minimum c'est d'abord chercher un point critique parce que pour que la fonction (continue) ait une minimum, cela signifie que la fonction va décroitre jusqu'à sa valeur minimale puis qu'elle va croitre après avoir atteint sa valeur minimale. Cela vient du faite que si elle avait continué à décroitre, alors elle n'aurait en réalité pas atteint son minimum. Lorsqu'elle décroit avant d'avoir atteint sa valeur minimale, sa dérivée (coefficient directeur de la tangente) est négative et après avoir atteint sa valeur minimale, lorsque la fonction croit, sa dérivée est positive. Comme la fonction f est supposée avoir une dérivée continue, alors on a que la dérivée vaut 0 pour le point critique (par le théorème des valeurs intermédiaires). Autrement dit, un minimiseur de la fonction f est nécessairement un point critique. Cependant, l'inverse n'est pas systématiquement vrai, une fois qu'on a déterminé qu'un point est un point critique, il faut déterminer s'il s'agit d'un minimum (graphique de gauche de la figure 3.7), un maximum (graphique du milieu) ou un point selle (graphique de droite).

Pour déterminer la nature d'un point critique, on étudie la dérivée seconde de la fonction f. Pour obtenir cette dérivée seconde f'', il suffit de dériver la dérivée f'. Une fois qu'on a f'', voici avec la proposition 3.2.4 la règle pour statuer quant à la nature du point critique.

Proposition 3.2.4 (Nature d'un point critique) *Pour une fonction f qui est deux fois dérivable, on a x_0 un point critique de f, alors*

- *si $f''(x_0) > 0$, alors x_0 est un minimum de f,*
- *si $f''(x_0) = 0$, alors x_0 est un point selle de f, et*
- *si $f''(x_0) < 0$, alors x_0 est un maximum de f.*

Exemple 3.2.5. *Soit $f(x) = (x - a)^2$. Voici les dérivées successives de la fonction f : $f'(x) = 2(x - a)$ et $f''(x) = 2$. Pour chercher les points critiques on chercher à annuler la dérivée première :*

$$f'(x) = 0 \Leftrightarrow 2(x - a) = 0 \Leftrightarrow x = a$$

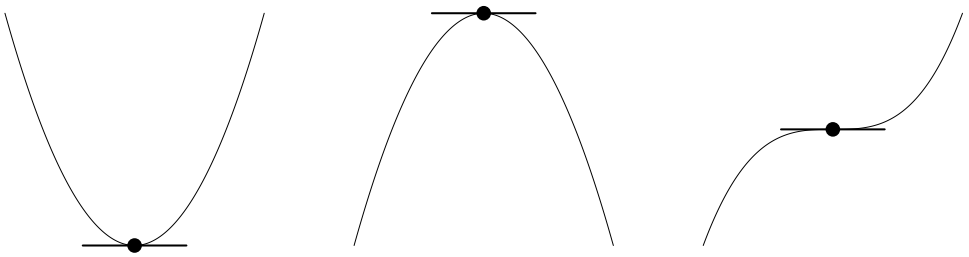


Figure 3.7 – *Les différents types de points critiques. Pour chacun des trois cas, le point correspond au point critique et la droite est la tangante à la courbe au point critique.*

et pour le point critique qu'on a trouvé $x_0 = a$, on a que $f''(x_0) = 2 > 0$. On a trouvé qu'il n'y avait qu'un seul point critique et qu'il s'agissait du minimum de la fonction.

Pour avoir une interprétation géométrique de la proposition 3.2.4, il faut comprendre que si la dérivée seconde $f''(x_0)$ est positive, cela veut dire que la dérivée première f' augmente autour de l'abscisse x_0. Si x_0 est un point critique (annulateur de la dérivée première), cela signifie que la dérivée première augmente en passant des valeurs négatives (avant x_0) à des valeurs positives (après x_0). Autrement dit, la fonction est décroissante (dérivée négative) avant x_0 et croissante (dérivée positive) après x_0. Le point critique x_0 est donc bien un minimum de la fonction f si $f''(x_0) > 0$. On peut avoir le même type de raisonnement pour les deux autres cas de la proposition 3.2.4.

3.2.3 Minimiser une fonction à deux variables

Lorsqu'on cherche à déterminer le minimum d'une fonction à deux variables, les outils sont légèrement différents de ceux utilisés pour le cas à une variable, de la section 3.2.2. Pour ce cas, la procédure reste la même :

- chercher les points critiques, et
- déterminer la nature des points critiques.

Voici ci-dessous les versions de la définition 3.2.3 et de la proposition 3.2.4 au cas de fonction à deux variables.

Définition 3.2.6 (Point critique d'une fonction à deux variables) *Pour une fonction $f(x, y)$ à deux variables, un point critique est un abscisse (x_0, y_0) qui vérifie :*

$$\frac{\partial f}{\partial x}(x_0, y_0) = 0 \quad et \quad \frac{\partial f}{\partial y}(x_0, y_0) = 0$$

à savoir c'est un annulateur des deux dérivées partielles de f.

Pour statuer sur la nature des points critiques d'une fonction à deux variables on introduit la matrice hessienne.

Définition 3.2.7 (Matrice hessienne) *Pour une fonction $f(x, y)$ à deux variables et pour un abscisse (x_0, y_0), sa matrice hessienne est :*

$$H(x_0, y_0) = \begin{pmatrix} \frac{\partial^2 f}{\partial x^2}(x_0, y_0) & \frac{\partial^2 f}{\partial x \partial y}(x_0, y_0) \\ \frac{\partial^2 f}{\partial x \partial y}(x_0, y_0) & \frac{\partial^2 f}{\partial y^2}(x_0, y_0) \end{pmatrix}$$

à savoir c'est la matrice des deux dérivées partielles secondes de f.

Définition 3.2.8 (Déterminant d'une matrice 2×2) *Pour une matrice de taille 2×2,*

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

son déterminant est $\det(M) = ad - cb$.

Définition 3.2.9 (Trace d'une matrice carrée) *Pour une matrice carré M de taille $n \times n$, la trace est la somme des éléments diagonaux :*

$$\text{trace}(M) = \sum_{i=1}^{n} M_{i,i},$$

où $M_{i,j}$ est la notation pour l'élément de la matrice M à la $i^{ème}$ ligne et la $j^{ème}$colonne.

Proposition 3.2.10 (Nature d'un point critique d'une fonction à deux variables) *Pour une fonction $f(x, y)$ a deux variables et un point critique (x_0, y_0), on a*

- *si $\det(H(x_0, y_0)) > 0$ et $\text{trace}(H(x_0, y_0)) > 0$, alors le point critique est un minimum,*
- *si $\det(H(x_0, y_0)) > 0$ et $\text{trace}(H(x_0, y_0)) < 0$, alors le point critique est un maximum,*
- *si $\det(H(x_0, y_0)) < 0$, alors le point critique est un point selle.*

S'il s'avère que $\det(H(x_0, y_0)) = 0$ alors il faut utiliser d'autres outils pour statuer sur la nature du point critique (hors programme).

Exemple 3.2.11. *Soit la fonction $f(x, y) = (x - 2y)^2 + (y - 2x)^2$. Les dérivées partielles sont :*

$$\frac{\partial f}{\partial x}(x, y) = 2(x - 2y) - 4(y - 2x) = 10x - 8y$$
$$\frac{\partial f}{\partial y}(x, y) = -4(x - 2y) + 2(y - 2x) = 10y - 8x$$

et les dérivées partielles secondes sont :

$$\frac{\partial^2 f}{\partial x^2}(x,y) = 10 \qquad \frac{\partial^2 f}{\partial y^2}(x,y) = 10 \qquad \frac{\partial^2 f}{\partial x \partial y}(x,y) = -8$$

Pour déterminer les points critiques, on résout le système suivant :

$$\begin{cases} \frac{\partial f}{\partial x}(x,y) &= 0 \\ \frac{\partial f}{\partial y}(x,y) &= 0 \end{cases} \Rightarrow \begin{cases} 10x - 8y &= 0 \\ 10y - 8x &= 0 \end{cases} \Rightarrow \begin{cases} x &= \frac{8}{10}y \\ y &= \frac{8}{10}x \end{cases} \Rightarrow \begin{cases} x &= \frac{8}{10}y \\ y &= \left(\frac{8}{10}\right)^2 y \end{cases} \Rightarrow \begin{cases} x & &= \frac{8}{10}y \\ y\left(1 - \left(\frac{8}{10}\right)^2\right) & &= 0 \end{cases} \Rightarrow \begin{cases} x &= 0 \\ y &= 0 \end{cases}$$

et on trouve que le seul point critique est $(0,0)$*. Pour ce point critique, la matrice hessienne est :*

$$H(0,0) = \begin{pmatrix} 10 & -8 \\ -8 & 10 \end{pmatrix}$$

dont le déterminant vaut $\det(H(0,0)) = 10 \times 10 - 8 \times 8 = 36$ *et dont la trace est* $\text{trace}(H(0,0)) = 10 + 10 = 20$*. Comme* $\det(H(0,0)) > 0$ *et* $\text{trace}(H(0,0)) > 0$*, alors le point critique* $(0,0)$ *est un minimum de la fonction* $f(x,y)$*.*

3.2.3.1 Minimiser le critère des moindres carrés

Dans cette section, on utilise les notions introduites dans les sections 3.2.2 et 3.2.3 pour déterminer le minimum du critère des moindres carrés. A savoir, il s'agit de trouver le couple de valeurs (μ, β) qui minimise la fonction $C(\mu, \beta)$, c'est-à-dire qui minimise les erreurs de prédiction du modèle de régression linéaire simple.

Pour cela, on commence par déterminer les dérivées partielles premières du critère C :

$$\begin{aligned}
\frac{\partial C}{\partial \mu}(\mu,\beta) &= \frac{\partial}{\partial \mu}\left(\frac{1}{n}\sum_{i=1}^{n}(y_i - (\mu + \beta x_i))^2\right) \\
&= \frac{1}{n}\frac{\partial}{\partial \mu}\left(\sum_{i=1}^{n}(y_i - (\mu + \beta x_i))^2\right) && \text{(puisque } (au)' = au' \text{ si } a \text{ est une constante)} \\
&= \frac{1}{n}\sum_{i=1}^{n}\frac{\partial}{\partial \mu}(y_i - (\mu + \beta x_i))^2 && \text{(puisque } (u+v)' = u' + v' \text{)} \\
&= \frac{1}{n}\sum_{i=1}^{n} 2(y_i - (\mu + \beta x_i)) \times \underbrace{\frac{\partial}{\partial \mu}(y_i - (\mu + \beta x_i))}_{=-1} && \text{(puisque } (u^2)' = 2uu' \text{)} \\
&= -2\frac{1}{n}\sum_{i=1}^{n}(y_i - (\mu + \beta x_i)) \\
&= -2\left(\frac{1}{n}\sum_{i=1}^{n} y_i - \frac{1}{n}\sum_{i=1}^{n}\mu - \frac{1}{n}\sum_{i=1}^{n}\beta x_i\right) \\
&= -2\left(\overline{y} - \mu - \beta\overline{x}\right) \\
\frac{\partial C}{\partial \beta}(\mu,\beta) &= \frac{\partial}{\partial \beta}\left(\frac{1}{n}\sum_{i=1}^{n}(y_i - (\mu + \beta x_i))^2\right) \\
&= \frac{1}{n}\sum_{i=1}^{n}\frac{\partial}{\partial \beta}(y_i - (\mu + \beta x_i))^2 \\
&= \frac{1}{n}\sum_{i=1}^{n} 2(y_i - (\mu + \beta x_i)) \times \underbrace{\frac{\partial}{\partial \beta}(y_i - (\mu + \beta x_i))}_{-x_i} \\
&= -2\frac{1}{n}\sum_{i=1}^{n} x_i(y_i - (\mu + \beta x_i)) \\
&= -2\left(\frac{1}{n}\sum_{i=1}^{n} x_i y_i - \frac{1}{n}\sum_{i=1}^{n}\mu x_i - \frac{1}{n}\sum_{i=1}^{n}\beta x_i^2\right) \\
&= -2\left(\overline{xy} - \mu\overline{x} - \beta\overline{x^2}\right)
\end{aligned}$$

En annulant les dérivées partielles on trouve le point critique avec les calculs suivants :

$$\begin{aligned}
\frac{\partial C}{\partial \mu}(\mu,\beta) = 0 &\Leftrightarrow -2\left(\overline{y} - \mu - \beta\overline{x}\right) = 0 \\
&\Leftrightarrow \overline{y} - \mu - \beta\overline{x} = 0 \\
&\Leftrightarrow \mu = \overline{y} - \beta\overline{x} \\
\frac{\partial C}{\partial \beta}(\mu,\beta) = 0 &\Leftrightarrow -2\left(\overline{xy} - \mu\overline{x} - \beta\overline{x^2}\right) = 0 \\
&\Leftrightarrow \overline{xy} - \left(\overline{y} - \beta\overline{x}\right)\overline{x} - \beta\overline{x^2} = 0 \\
&\Leftrightarrow \overline{xy} - \overline{x}\,\overline{y} + \beta\overline{x}^2 - \beta\overline{x^2} = 0 \\
&\Leftrightarrow \beta\left(\overline{x^2} - \overline{x}^2\right) = \overline{xy} - \overline{x}\,\overline{y} \\
&\Leftrightarrow \beta = \frac{\overline{xy} - \overline{x}\,\overline{y}}{\overline{x^2} - \overline{x}^2}
\end{aligned}$$

Le point critique est donc le suivant :

$$(\mu_0, \beta_0) = \left(\overline{y} - \beta_0\overline{x} \quad , \quad \frac{\overline{xy} - \overline{x}\,\overline{y}}{\overline{x^2} - \overline{x}^2}\right)$$

et il reste à montrer si ce point critique est bien un minimum du critère des moindres carrés. Pour cela, on calcule les dérivées partielles secondes :

$$\frac{\partial^2 C}{\partial \mu^2}(\mu,\beta) = 2 \qquad \frac{\partial^2 C}{\partial \beta^2}(\mu,\beta) = 2\overline{x^2} \qquad \frac{\partial^2 C}{\partial \mu \partial \beta}(\mu,\beta) = 2\overline{x}$$

et donc la matrice hessienne est :

$$H(\mu,\beta) = \begin{pmatrix} 2 & 2\overline{x} \\ 2\overline{x} & 2\overline{x^2} \end{pmatrix}$$

dont le déterminant et la trace, pour le point critique, sont :

$$\det(H(\mu_0,\beta_0)) = 4\left(\overline{x^2} - \overline{x}^2\right) = 4s^2 \qquad \text{et} \qquad \operatorname{trace}(H(\mu_0,\beta_0)) = 2 + 2\overline{x^2}$$

Comme $s^2 > 0$ et $\overline{x^2} > 0$, on a que le déterminant et la trace de la matrice hessienne du point critique sont positifs. Autrement dit, le point critique est bien un minimum du critère des moindres carrés.

Un raisonnement supplémentaire, qui n'est pas détaillé ici, permet de justifier qu'il s'agisse de l'unique minimum de cette fonction, ce qui nous permet d'établir la proposition suivante.

Proposition 3.2.12 (Minimum du critère des moindres carrés) *Pour le modèle de régression linéaire simple, pour*

$$\mu_0 = \overline{y} - \beta_0\overline{x} \qquad \textit{et} \qquad \beta_0 = \frac{\overline{xy} - \overline{x}\,\overline{y}}{\overline{x^2} - \overline{x}^2},$$

le critère des moindres carrés atteint sont unique minimum.

Ceci permet de justifier que les estimateurs obtenus ont les expressions données par la proposition 3.2.12. Par convention, on note ce minimum du critère des moindres carrés comme des estimateurs :

$$\hat{\mu} = \overline{y} - \hat{\beta}\overline{x} \qquad \text{et} \qquad \hat{\beta} = \frac{\overline{xy} - \overline{x}\,\overline{y}}{\overline{x^2} - \overline{x}^2},$$

et ils sont appelés "estimateurs des moindres carrés".

Mise en pratique des notions de la section 3.2

Exercice 3.2.1 (Minimum et maximum)**.** Déterminez les minimums et maximums des fonctions suivantes :

1. $f(x) = 3x^2 + x$,
2. $f(x) = \exp\{-x^2\}$,
3. $f(x) = (x-4)^3$

Exercice 3.2.2 (Minimisation d'un critère des moindres carrés)**.** Pour cet exercice on considère le modèle suivant :

$$y_i = \beta x_i + \varepsilon_i$$

pour lequel une prédiction aurait pour expression : $\hat{y}_i = \hat{\beta} x_i$.

1. Ecrivez le crtière des moindres carrés pour ce modèle :

$$C(\beta) = \dots$$

2. Déterminez le minimum de ce critère des moindres carrés.

3.3 Qualité d'ajustement

Comme vu en section 3.2, les estimateurs des moindres carrés correspondent aux meilleures valeurs de μ et β afin d'obtenir un modèle de régression linéaire simple fournissant les meilleures prédictions. Cependant, en pratique lorsqu'on dispose d'un jeu de données, cela ne garantit pas que les prédictions seront suffisamment bonnes (seulement les meilleures avec ce modèle). Pour se donner une idée de à quel point le modèle est correct pour décrire les données, et donc que les prédictions $\hat{y}_i$ sont proches des vraies données y_i, on peut utiliser un indicateurs numérique : le coefficient de détermination.

Avant d'introduire ce coefficient, on introduit le théorème 3.3.1 qui donne la formule de la décomposition de la variance.

Théorème 3.3.1 (Décomposition de la variance) *Soit $y = (y_1, \dots, y_n)$ des données quantitatives, et $\hat{y} = (\hat{y}_1, \dots, \hat{y}_n)$ les prédictions du modèle de régression linéaire simple, alors le facteur principal de la variance empirique peut se décomposer de la manière suivante :*

$$\sum_{i=1}^{n}(y_i - \overline{y})^2 = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2$$

Pour la suite, on note de la manière suivante chacun des termes de la décomposition de la variance :

- SCT $= \sum_{i=1}^{n}(y_i - \overline{y})^2$ est la Somme des Carrés Totale, qui correspond à un terme de la variance empirique s_y^2.
- SCR $= \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ est la Somme des Carrés Résiduels, qui correspond à un terme du critère des moindres carrés, et à un terme de la variance empirique des résidus.
- SCM $= \sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2$ est la Somme des Carrés du Modèle, qui correspond à un terme de la variance empirique des prédictions $s_{\hat{y}}^2$ (puisque la moyenne des prédictions est égale à la moyenne des données y_i).

Autrement dit, on peut réécrire la décomposition de la variance de la manière suivante :

$$\text{SCT} = \text{SCR} + \text{SCM}$$

Pour interprétation, la variance empirique des données se décompose comme la variance des prédiction du modèle plus la variance des résidus. Cette décomposition et son interprétation permettent d'introduire le coefficient de détermination.

Définition 3.3.2 (Coefficient de détermination) *Le coefficient de détermination est noté R^2 et son expression est :*

$$R^2 = \frac{\text{SCM}}{\text{SCT}}$$

Propriété 3.3.3 (Interprétation du coefficient de détermination) *Ce coefficient est entre 0 et 1 et :*

- *lorsqu'il est proche de 1, il y a vraisemblablement une liaison linéaire entre les deux variables, et*
- *lorsqu'il est proche de 0, il n'y a vraisemblablement pas une liaison linéaire entre les deux variables.*

En plus de cette interprétation, ce coefficient de détermination a un lien avec certaines quantités en lien avec la potentielle liaison entre les deux variables.

Proposition 3.3.4 (Coefficient de détermination et corrélation linéaire) *Pour R^2, le coefficient de détermination, et $r(x, y)$, le coefficient de corrélation, on a :*

$$R^2 = r(x, y)^2.$$

Proposition 3.3.5 (Coefficient de détermination et estimation du coefficient de pente) *Pour R^2, le coefficient de détermination, et $\hat{\beta}$, l'estimation du coefficient de pente, on a :*

$$R^2 = \hat{\beta}^2 \frac{s_x^2}{s_y^2}.$$

Avec les propositions 3.3.4 et 3.3.5, on peut interpréter le coefficient de détermination comme une intensité de liaison, peu importe que la liaison soit négative ou positive. Il s'agit donc bien de quantifier à quel point la relation est linéaire entre les deux variables.

Pour compléter cette interprétation, de manière assez standard on détermine que R^2 correspond à un pourcentage de variabilité expliquée par le modèle. Par exemple, si $R^2 = 0.75$, on interprète que parmi les variations aléatoires des données y, il y a une proportion de 75% de cette variabilité qu'on peut déterminer et expliquer à l'aide d'une liaison linéaire entre X et Y. Cette interprétation peut se comprendre grace à la formule de la définition 3.3.2, pour laquelle :

- SCT, correspond (à un facteur près) à la variance des données y, et
- SCM, correspond (à un facteur près) à la variance expliquée par le modèle.

Et donc, le ratio entre les deux correspond à la proportion de variance expliquée par le modèle.

Mise en pratique des notions de la section 3.3

Exercice 3.3.1 (Démonstration de la décomposition de la variance)**.** Le but de cet exercice est de faire la démonstration du théorème 3.3.1, et pour cela on va commencer par démontrer le lemme 3.3.6 qui sera utile pour cette démonstration.

Lemme 3.3.6 (Produit scalaire des residus et des données x_i) *Dans le cadre du modèle de régression linéaire simple, on a*

$$\sum_{i=1}^{n} e_i x_i = 0.$$

1. Pour démontrer le lemme 3.3.6, posez le calcul $\sum_{i=1}^n e_i x_i$ en remplaçant e_i par son expression, puis en remplaçant au fur et à mesure $\hat{\mu}$ par son expression et $\hat{\beta}$ par son expression.
2. Pour démontrer la décomposition de la variance, commencez par expliquer que $(y_i - \overline{y})^2 = \big((y_i - \hat{y}_i) + (\hat{y}_i - \overline{y})\big)^2$.
3. Utilisez l'identité remarquable $(a+b)^2 = a^2 + b^2 + 2ab$ avec $a = y_i - \hat{y}_i$ et $b = \hat{y}_i - \overline{y}$.
4. Distribuez les opérateurs de somme et montrez qu'on obtient :

$$\sum_{i=1}^{n} (y_i - \overline{y})^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 + \sum_{i=1}^{n} (\hat{y}_i - \overline{y})^2 + 2\sum_{i=1}^{n} e_i (\hat{y}_i - \overline{y})$$

5. Développez le produit dans le terme $\sum_{i=1}^n e_i(\hat{y}_i - \overline{y})$ et remplacez $\hat{y}_i$ par son expression. De plus distribuez l'opérateur de somme pour faire apparaître les termes $\sum_{i=1}^n e_i$ et $\sum_{i=1}^n e_i x_i$.
6. Utilisez le lemme 3.3.6 et la proposition 3.1.9 pour montrer que $\sum_{i=1}^n e_i(\hat{y}_i - \overline{y}) = 0$.
7. En déduire le résultat de la décomposition de la variance.

Exercice 3.3.2 (Application)**.** Avec les données de l'exercice 3.1.1, calculez le coefficient de détermination avec la formule de la décomposition de la variance, et dans un second temps avec la formule de la proposition 3.3.5.

3.4 Diapos de cours et exercices de travaux dirigés

Chap. 2 – **Ajustement linéaire**

Motivations et objectifs

Contexte
On dispose de mesures d'individus pour deux critères quantitatifs. Pour le $i^{\text{ème}}$ individu statistique, les mesures sont notées x_i et y_i.

Motivations
1. Identifier le lien entre les deux variables.
2. Compléter des données manquantes.
3. Prédire de nouvelles données.

Approche
Tracez une courbe qui respecte la forme du nuage de points.
La courbe ne doit pas passer exactement par chaque point. Les observations sont considérées comme des variations autour d'un modèle inconnu.

$$\textbf{données} = \textbf{modèle} + \text{variation}$$

Objectif de ce chapitre :
Comment on obtient l'équation de la droite ?

Couple de variables

Notations :
Soient X et Y deux variables quantitatives. On s'intéresse au couple des variables (X, Y) dont les "valeurs" possibles sont les paires (x_i, y_i) où x_i et y_i sont respectivement des mesures de X et Y.

Nuage de points :
Le nuage des points est un graphique contenant des points de coordonnées : $M_i : (x_i, y_i)$.

.
.
.
.
.
.
.
.
.
.
.
.

Liaison entre les deux variables :
Au regard du nuage de points, à l'œil nu, on peut déceler une liaison particulière à partir de la forme du nuage. Il est cependant nécessaire de le faire mathématiquement.

Corrélation linéaire

Rappel : Corrélation
La corrélation linéaire empirique est donnée par la formule suivante :

$$r(x,y) = \frac{\text{cov}(x,y)}{s_x \times s_y} \quad \text{où} \quad \text{cov}(x,y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \overline{xy} - \bar{x}\bar{y},$$

où $\overline{xy} = \frac{1}{n}\sum_{i=1}^{n} x_i y_i$.
Cet indicateur numérique permet de quantifier une relation linéaire entre des données $x_1, \ldots, x_n$ et $y_1, \ldots, y_n$.

.
.
.
.
.

Ce n'est pas la seule liaison possible, mais c'est la plus simple à mettre à lumière.

Régression linéaire : rappels

Régression linéaire :
Pour des données $x_1, \ldots, x_n$ et $y_1, \ldots, y_n$, la régression linéaire est un modèle qui permet de prédire la donnée y_i à partir de x_i. L'équation de ce modèle est :

$$y_i = \mu + \beta x_i + \varepsilon_i$$

où μ et β sont des paramètres inconnus à estimer, et ε_i est l'erreur du modèle.

Modèle ajusté et corrélation :
On parle de modèle **ajusté** pour l'équation du modèle :

$$y_i = \hat{\mu} + \hat{\beta} x_i + e_i$$

où e_i est l'erreur de prédiction pour la $i^{\text{ème}}$ donnée. Les estimations des paramètres sont données par :

$$\hat{\beta} = \frac{\text{cov}(x,y)}{s_x^2} = r(x,y) \times \frac{s_x}{s_y} \quad \text{et} \quad \hat{\mu} = \bar{y} - \hat{\beta}\bar{x}.$$

Prédiction :
Pour ce modèle de régression linéaire, la prédiction de la valeur de y_i pour une mesure x_i donnée est :

$$\hat{y} = \hat{\mu} + \hat{\beta} x_i.$$

Autres rappels

Résidus :
Le résidu e_i est l'erreur de prédiction du modèle. Le résidu est donné par : $e_i = \hat{y}_i - y_i$.

Somme des carrés :
- La somme des carrés totale (SCT) est : $\text{SCT} = \sum_{i=1}^{n}(y_i - \bar{y})^2$
- La somme des carrés du modèle (SCM) est : $\text{SCM} = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$
- La somme des carrés résiduels (SCR) est : $\text{SCR} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$

Décomposition de la variance :
La somme des carrés totale, qui est relatif à la variance, se décompose de la manière suivante : $\text{SCT} = \text{SCM} + \text{SCR}$.

Coefficient de détermination :
Le coefficient de détermination R^2 est donné par : $R^2 = \text{SCM} \,/\, \text{SCT}$.
Ce coefficient est entre 0 et 1.
- Lorsqu'il est proche de 1, il y a vraisemblablement une liaison linéaire entre les deux variables.
- Lorsqu'il est proche de 0, il n'y a vraisemblablement pas une liaison linéaire entre les deux variables.

Méthode des moindres carrés

Estimation :
Jusque là, les estimateurs des paramètres μ et β ont été donné. Dans ce qui suit, il est expliqué comment on obtient ces estimateurs.

L'intuition pour trouver ces estimateurs est de chercher à minimiser les erreurs de prédictions. En particulier, on va minimiser la distance (au carré) entre les prédictions $\hat{y}_i$ et les données y_i, c'est la **méthode des moindres carrés**. C'est la méthode qu'il sera utilisé pour estimer les paramètres d'un modèle, pour une grande partie des modèles possibles.

Méthode des moindres carrés :
La distance entre les prédictions $\hat{y} = (\hat{y}_1, \ldots, \hat{y}_n)$ et les données $y = (y_1, \ldots, y_n)$ est donnée par :

$$d(\hat{y}, y) = \sqrt{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} = \sqrt{\sum_{i=1}^{n}\left(y_i - (\hat{\mu} + \hat{\beta}x_i)\right)^2} = \sqrt{\sum_{i=1}^{n} e_i^2}$$

Pour obtenir cette formule, il faut déterminer les valeurs de μ et β.
Pour cela, la méthode des moindres carrés va consister à chercher les valeurs μ et β qui minimise le carré de cette distance :

$$\sum_{i=1}^{n}(y_i - (\mu + \beta x_i))^2$$

Pour la suite, il va donc falloir minimiser une fonction par rapport à deux variables : μ et β.

Minimiser une fonction

Fonction à une variable réelle :
Soit f une fonction dont on souhaite trouver la valeur x_0 telle que $\forall x \in \mathbb{R}, f(x_0) \leq f(x)$ (x_0 est le minimiseur de la fonction f).

Pour cela, on calcule la dérivée f' de la fonction f et on cherche à résoudre l'équation $f'(x) = 0$. Les solutions de cette équation sont les points critiques qui sont des candidats à être des minimiseurs de la fonction f.

.
.
.
.
.
.
.
.
.
.
.
.
.
.

La fonction qu'on cherche à miniser est f :

$$f(x) = x^4 + \frac{3}{2}(x-4)^3 - 72(x-3)$$

La dérivée f' de cette fonction est :

$$f'(x) = 4x^3 + \frac{9}{2}(x-4)^2 - 72$$

Après quelques calculs on trouve 3 solutions à l'équation $f'(x) = 0$:
$x_1 = -3.55$, $x_2 = 0$ et $x_3 = 2.55$.

La fonction est minimisée en $x = -3.55$.

Minimiser une fonction (2)

Fonction à deux variables réelles :
Soit f une fonction à deux variables. Pour chaque couple (x, y), la fonction f renvoie une valeur $z : f(x, y) = z$. On souhaite trouver le couple (x_0, y_0) tel que $\forall (x, y) \in \mathbb{R}^2, f(x_0, y_0) \leq f(x, y)$ ((x_0, y_0) est le couple minimiseur de la fonction f).

Pour cela, on calcule les **dérivées partielles** de la fonction f, les deux dérivées partielles sont notées : $\frac{\partial f}{\partial x}$ et $\frac{\partial f}{\partial y}$. On résout ensuite les équations des dérivées partielles :

$$\frac{\partial f}{\partial x}(x, y) = 0 \quad \text{et} \quad \frac{\partial f}{\partial y}(x, y) = 0.$$

Dérivées partielles :
La dérivée partielle $\frac{\partial f}{\partial x}$ est obtenue en dérivant la fonction $f(x, y)$ en considérant y comme une constante.
De la même manière, pour la dérivée partielle $\frac{\partial f}{\partial y}$, x est considérée comme une constante.

Exemple :
Soit la fonction à deux variables : $f(x, y) = x^2 + 4xy + y^3$. Les dérivées partielles sont :

$$\frac{\partial f}{\partial x}(x, y) = 2x + 4y \quad \text{et} \quad \frac{\partial f}{\partial y}(x, y) = 4x + 3y^2$$

Ces deux dérivées partielles sont égales à 0 pour les couples :
$(x, y) = (0, 0)$ et $(x, y) = \left(-\frac{4}{3}, \frac{4}{3}\right)$
Ces couples sont les points critiques de f. Il faudrait faire un peu plus de calculs pour savoir si l'un d'eux est le minimiseur.

Moindres carrés et régression linéaire

A minimiser :
Pour l'application de la méthode des moindres carrés, on cherche μ et β qui minimise la fonction introduite précédemment, qu'on va noter C :

$$C(\mu, \beta) = \sum_{i=1}^{n} (y_i - (\mu + \beta x_i))^2$$

Les dérivées partielles de C sont :

$$\begin{aligned}\frac{\partial C}{\partial \mu}(\mu, \beta) &= \sum_{i=1}^{n} 2 \times \frac{\partial (y_i - (\mu + \beta x_i))}{\partial \mu} (y_i - (\mu + \beta x_i)) \\ &= \sum_{i=1}^{n} 2 \times (-1) \times (y_i - (\mu + \beta x_i)) \\ &= -2 \sum_{i=1}^{n} (y_i - (\mu + \beta x_i))\end{aligned}$$

Moindres carrés et régression linéaire

A minimiser :
Pour l'application de la méthode des moindres carrés, on cherche μ et β qui minimise la fonction introduite précédemment, qu'on va noter C :

$$C(\mu, \beta) = \sum_{i=1}^{n} (y_i - (\mu + \beta x_i))^2$$

Les dérivées partielles de C sont :

$$\begin{aligned}\frac{\partial C}{\partial \beta}(\mu, \beta) &= \sum_{i=1}^{n} 2 \times \frac{\partial (y_i - (\mu + \beta x_i))}{\partial \beta} (y_i - (\mu + \beta x_i)) \\ &= \sum_{i=1}^{n} 2 \times (-x_i) \times (y_i - (\mu + \beta x_i)) \\ &= -2 \sum_{i=1}^{n} x_i (y_i - (\mu + \beta x_i))\end{aligned}$$

Moindres carrés et régression linéaire

Les dérivées partielles de C sont :

$$\frac{\partial C}{\partial \mu}(\mu, \beta)) = -2 \sum_{i=1}^{n} (y_i - (\mu + \beta x_i)) \quad \text{et} \quad \frac{\partial C}{\partial \beta}(\mu, \beta) = -2 \sum_{i=1}^{n} x_i (y_i - (\mu + \beta x$$

On cherche maintenant à résoudre le système d'équations :

$$\frac{\partial C}{\partial \mu}(\mu, \beta)) = 0 \quad \text{et} \quad \frac{\partial C}{\partial \beta}(\mu, \beta)) = 0.$$

$$-2 \sum_{i=1}^{n} \left(y_i - (\mu + \beta x_i)\right) = 0$$

$$\sum_{i=1}^{n} y_i - \sum_{i=1}^{n} \mu - \beta \sum_{i=1}^{n} x_i = 0$$

$$\frac{1}{n} \sum_{i=1}^{n} y_i - \frac{1}{n} \sum_{i=1}^{n} \mu - \beta \frac{1}{n} \sum_{i=1}^{n} x_i = \frac{1}{n} \times 0$$

$$\bar{y} - \mu - \beta \bar{x} = 0$$

$$\mu = \bar{y} - \beta \bar{x} \quad (1)$$

$$\frac{\partial C}{\partial \beta}(\mu, \beta)) = 0$$

$$-2 \sum_{i=1}^{n} x_i \left(y_i - (\mu + \beta x_i)\right) = 0$$

$$\sum_{i=1}^{n} x_i y_i - \mu \sum_{i=1}^{n} x_i - \beta \sum_{i=1}^{n} x_i^2 = 0$$

$$\frac{1}{n} \sum_{i=1}^{n} x_i y_i - \mu \frac{1}{n} \sum_{i=1}^{n} x_i - \beta \frac{1}{n} \sum_{i=1}^{n} x_i^2 = \frac{1}{n} \times 0$$

$$\overline{xy} - \mu \bar{x} - \beta \overline{xx} = 0$$

$$\beta = \frac{\overline{xy} - \mu \bar{x}}{\overline{xx}} \quad (2)$$

Moindres carrés et régression linéaire

Les dérivées partielles de C sont :

$$\frac{\partial C}{\partial \mu}(\mu,\beta)) = -2\sum_{i=1}^{n}(y_i - (\mu+\beta x_i)) \quad \text{et} \quad \frac{\partial C}{\partial \beta}(\mu,\beta) = -2\sum_{i=1}^{n} x_i\,(y_i - (\mu+\beta x_i))$$

En cherchant à résoudre les équations des dérivées partielles on a obtenu :

$$(1) \qquad \mu = \bar{y} - \beta\bar{x}$$

$$(2) \qquad \beta = \frac{\overline{xy} - \mu\bar{x}}{\overline{xx}}$$

En utilisant (1) dans l'équation (2), on obtient :

$$\begin{aligned}
\beta &= \frac{\overline{xy} - (\bar{y} - \beta\bar{x})\bar{x}}{\overline{xx}} \\
\overline{xx}\beta &= \overline{xy} - \bar{y}\bar{x} + \beta\bar{x}^2 \\
\overline{xx}\beta - \beta\bar{x}^2 &= \overline{xy} - \bar{y}\bar{x} \\
\beta(\overline{xx} - \bar{x}^2) &= \overline{xy} - \bar{y}\bar{x} \\
\beta &= \frac{\overline{xy} - \bar{y}\bar{x}}{\overline{xx} - \bar{x}^2} \\
\beta &= \frac{\mathrm{cov}(x,y)}{s_x^2}
\end{aligned}$$

Au final, on obtient :

$$\mu = \bar{y} - \beta\bar{x} \quad \text{et} \quad \beta = \frac{\mathrm{cov}(x,y)}{s_x^2}.$$

Moindres carrés et régression linéaire

A minimiser :
Pour l'application de la méthode des moindres carrés, on cherche μ et β qui minimise la fonction introduite précédemment, qu'on va noter C :

$$C(\mu,\beta) = \sum_{i=1}^{n}(y_i - (\mu+\beta x_i))^2$$

Les dérivées partielles de C sont :

$$\frac{\partial C}{\partial \mu}(\mu,\beta)) = -2\sum_{i=1}^{n}(y_i - (\mu+\beta x_i)) \quad \text{et} \quad \frac{\partial C}{\partial \beta}(\mu,\beta) = -2\sum_{i=1}^{n} x_i\,(y_i - (\mu+\beta x_i))$$

En cherchant à résoudre les équations des dérivées partielles on a obtenu les point critique :

$$\mu = \bar{y} - \beta\bar{x} \quad \text{et} \quad \beta = \frac{\mathrm{cov}(x,y)}{s_x^2}.$$

Pour cet exemple-ci, il n'est pas nécessaire de montrer que ce point critique est le minimiseur. L'argument à utiliser est que la fonction qu'on cherche à minimiser est une parabole (en deux dimensions) de la forme x^2.

On obtient donc les estimateurs $\hat{\mu}$ et $\hat{\beta}$ donnés par :

$$\hat{\mu} = \bar{y} - \hat{\beta}\bar{x} \quad \text{et} \quad \hat{\beta} = \frac{\mathrm{cov}(x,y)}{s_x^2}.$$

Résumé

Les éléments dont il faut se souvenir :

- Modèle de régression linéaire
- Comment déterminer les estimateurs $\hat{\mu}$ et $\hat{\beta}$?
- Méthode des moindres carrés
- Minimiser une fonction (recherche de points critiques)
- Retrouver les estimateurs à partir de la méthode des moindres carrés

TD 1 – Optimisation et moindres carrés

Mémo sur l'optimisation

Comme trouver un point critique d'une fonction $f(x)$:

Pour trouver un point critique d'une fonction f à une seule variable, il suffit de calculer la dérivée f' et de déterminer la ou les solutions de l'équation $f'(x) = 0$. Chaque solution de cette équation est un point critique de f.

Comment trouver un maximum local d'une fonction $f(x)$:

Soient x_0 un point critique de f et f'' sa dérivée seconde. Si $f''(x_0) < 0$, alors ce point critique est un maximum local.

Trouver un minimum local d'une fonction $f(x)$:

Soient x_0 un point critique de f et f'' sa dérivée seconde. Si $f''(x_0) > 0$, alors ce point critique est un minimum local.

Trouver un point critique local d'une fonction $f(x, y)$:

Pour trouver un point critique d'une fonction f à deux variables, il suffit de calculer les dérivées partielles $\frac{\partial f}{\partial x}$ et $\frac{\partial f}{\partial y}$ et de déterminer le ou les couples (x_0, y_0) qui vérifient les deux équations $\frac{\partial f}{\partial x}(x_0, y_0) = 0$ et $\frac{\partial f}{\partial y}(x_0, y_0) = 0$. Chaque couple vérifiant ces deux équations est un point critique de f.

Qu'est-ce qu'un maximum global :

Un maximum global est un point x_0 pour lequel $f(x_0) \geq f(x)$ quelque soit $x \in \mathbb{R}$.

C'est un point x_0 pour lequel la fonction f est maximale.

Qu'est-ce qu'un maximum local :

Un maximum local est un point x_0 pour lequel il existe un intervalle V, contenant x_0, pour lequel $f(x_0) \geq f(x)$ pour tout $x \in V$.

C'est un point x_0 qui maximise la fonction f localement (autour de x_0). Autrement dit, la fonction f peut atteindre des valeurs plus grandes que $f(x_0)$, mais pour un point x_1 proche de x_0, la valeur $f(x_1)$ sera plus petite que $f(x_0)$.

Minimum local et global :

Pour obtenir les définitions de minimum local et global, il suffit de remplacer le signe "$\geq$" par "$\leq$" dans les définitions précédentes concernant le maximum.

Différentre entre optimum (minimum ou maximum) local et un optimum global :

Un optimum global est un optimum local mais l'inverse n'est pas forcément vrai.

Exercice 1. Soit $f(x) = x^2$ pour $x \in \mathbb{R}$.

1. Montrez que $x = 0$ est un point critique de f.
2. Déterminez si ce point critique est un minimum local ou un maximum local.
3. Argumentez comme vous pouvez pour expliquer s'il s'agit d'un optimum local ou global.

Exercice 2. Soit $f(x) = x^3$ pour $x \in \mathbb{R}$.

1. Déterminez le seul point critique de f.
2. Montrez qu'il ne s'agit pas d'un optimum local. Vous tracez le graphique de la fonction pour appuyer votre argumentation.

Exercice 3. Soit $f(x) = -(x-2)^2(x+2)^2$ pour $x \in \mathbb{R}$.

1. Montrez que $f'(x) = -4x(x-2)(x+2)$.
2. En déduire les points critiques de f.
3. Déterminez la dérivée seconde f''.
4. Déterminez si les points critiques de f sont des minimums ou des maximums locaux.
5. Expliquez, sans vous appuyer sur le graphique de la fonction f, pourquoi il n'y a pas de minimum global pour la fonction f.

Exercice 4. Soit $f(x) = (x-2)^2 + (x+5)^2$ pour $x \in \mathbb{R}$.

1. Déterminez le point critique de f.
2. Montrez qu'il s'agit d'un minimum local.
3. Argumentez comme vous pouvez pour expliquer qu'il s'agit d'un minimum global.

Exercice 5. Soit $a = (a_1, \ldots, a_n)$, une suite de nombres réels. On note $\overline{a}$, la moyenne de cette suite de nombre. Soit $f(x) = \sum_{i=1}^{n}(x-a_i)^2$.

1. Expliquez pourquoi $f'(x) = \sum_{i=1}^{n}\left((x-a_i)^2\right)'$.
2. En déduire que $f'(x) = 2n(x-\overline{a})$.
3. En déduire le point critique de f.
4. Déterminez la dérivée seconde $f''(x)$.
5. Déterminez si le point critique est un minimum local ou un maximum local.

Exercice 6. Soit $f(x,y) = x^2 + y^2$ pour $(x,y) \in \mathbb{R}^2$.

1. Calculez les dérivées partielles de f.
2. Montrez que le point critique de f est $(x,y) = (0,0)$.
3. Calculez $f(0,0)$ et argumentez comme vous pouvez pour expliquez si le point critique est un minimum global, un maximum global ou aucun des deux.

Exercice 7. Soit $f(x,y) = 2x^2 + (y-1)^2$ pour $(x,y) \in \mathbb{R}^2$.

1. Calculez les dérivées partielles de f.
2. Déterminez le point critique.
3. Argumentez comme vous pouvez pour expliquer si le point critique est un minimum global, un maximum global ou aucun des deux.

Exercice 8. Soit $f(x,y) = (x-y)^2 + (x-1)^2$ pour $(x,y) \in \mathbb{R}^2$.

1. Calculez les dérivées partielles de f.
2. Détezminez le point critique.
3. Argumentez comme vous pouvez pour expliquer si le point critique est un minimum global, un maximum global ou aucun des deux.

Exercice 9. Dans un contexte de régression linéaire, on s'intéresse à prédire des données $y = (y_1, \ldots, y_n)$ à partir de données transformées $f(x) = \big(f(x_1), \ldots, f(x_n)\big)$. Soit C le critère des moindres carrés qu'on cherche à minimiser en trouvant les bonnes valeurs de μ et de β :

$$C(\mu, \beta) = \sum_{i=1}^{n} \left(y_i - (\mu + \beta f(x_i))\right)^2 .$$

Pour cet exercice, on utiliser les notations suivantes :
$\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$,

$\overline{f(x)} = \frac{1}{n}\sum_{i=1}^{n} f(x_i)$,
$\overline{yf(x)} = \frac{1}{n}\sum_{i=1}^{n} y_i f(x_i)$ et
$\overline{f(x)^2} = \frac{1}{n}\sum_{i=1}^{n} f(x_i)^2$.

1. Montrez que les dérivées partielles de f sont :

$$\frac{\partial C}{\partial \mu}(\mu,\beta) = -2n\overline{y} + 2n\mu + 2n\beta\overline{f(x)} \quad \text{et} \quad \frac{\partial C}{\partial \beta}(\mu,\beta) = -2n\overline{yf(x)} + 2n\mu\overline{f(x)} + 2n\beta\overline{f(x)^2}.$$

2. On note (μ_1, β_1) le point critique de C. Montrez que μ_1 doit vérifier $\mu_1 = \overline{y} - \beta_1\overline{f(x)}$.
3. Montrez de plus que β_1 doit vérifier :
$$\beta_1 = \frac{\overline{yf(x)} - \mu_1\overline{f(x)}}{\overline{f(x)^2}}.$$
4. Utilisez la liaison entre μ_1 et β_1 donnée à la question 2. pour montrer que :
$$\beta_1 = \frac{\overline{yf(x)} - \overline{y} \times \overline{f(x)}}{\overline{f(x)^2} - \overline{f(x)} \times \overline{f(x)}}.$$

Exercice 10. Dans un contexte de régression linéaire, on s'intéresse à prédire les données transformées $g(y) = (g(y_1), \ldots, g(y_n))$ à partir de données transformées $f(x) = (f(x_1), \ldots, f(x_n))$. Soit C le critère des moindres carrés qu'on cherche à minimiser en trouvant les bonnes valeurs de μ et de β :

$$C(\mu,\beta) = \sum_{i=1}^{n} \left(g(y_i) - (\mu + \beta f(x_i))\right)^2.$$

En suivant la même procédure que pour celle de l'exercice 9, montrez que le point critique de C est :

$$(\mu_1, \beta_1) = \left(\overline{g(y)} - \beta_1\overline{f(x)}\,,\ \frac{\overline{g(y)f(x)} - \overline{g(y)} \times \overline{f(x)}}{\overline{f(x)^2} - \overline{f(x)} \times \overline{f(x)}}\right)$$

CHAPITRE 4

Ajustement non-linéaire

Dans le chapitre 3, le modèle introduit permet d'analyser la liaison entre les variables X et Y, en ne pouvant la détecter que si la liaison est linéaire. Or en pratique, il y a des contextes pour lesquels cette approche n'est pas adapté, à savoir lorsqu'il y a une liaison non-linéaire entre les deux variables. Soit parce qu'il s'agit d'une liaison non-linéaire pour laquelle une approximation linéaire serait trop mauvaise, soit parce qu'avec les outils de linéarité on pourrait croire qu'il n'y a pas de liaison du tout alors qu'il y en a une.

Les graphiques de la figure 4.1 illustrent différentes situations de l'application d'une approche linéaire lorsque la liaison est en réalité non-linéaire. Sur le graphique de gauche, on constate que la droite de régression est horizontale. Autrement dit, l'estimation du coefficient β vaut 0, et de même, $R^2 = 0$. Ces résultats mènent à conclure qu'il n'y pas de liaison linéaire entre X et Y. Dans ce cas-là, il ne faut pas penser que cela induit qu'il n'y a pas de liaison, seulement parce qu'on ne détecte pas de liaison linéaire. En effet, sur cet exemple, on peut observer une liaison quadratique, autrement dit, il y a bien une liaison entre les données x et les données y.
Pour le graphique du milieu, on a une liaison non-linéaire (cubique) mais on peut se rendre compte qu'en utilisant une approche linéaire, cela permet d'avoir une description relativement fidèle du nuage de points. Autrement dit, la liaison linéaire semble ici être une approximation acceptable de la réelle liaison non-linéaire. Dans ce cas, on pourra préférer une approche linéaire, de part sa simplicité et pusiqu'elle fournit une bonne approximation de la réelle liaison.
Le graphique de droite est un autre type de liaison non-linéaire (qui s'appelle logistique), pour laquelle l'utilisation du modèle de régression linéaire n'est pas une approche adaptée. Non seulement, on commet des erreurs de prédiction, mais en plus à en croire les résultats de l'approche linéaire, on aurait tendance à supposer que pour des valeurs de X très grandes (en positif ou en négatif), la valeur de Y devrait aussi être très grande (en positif ou en négatif, respectivement par rapport au signe de X). Cependant, avec le nuage de point, on constate que les valeurs de Y stagne dès lors que X prend des valeurs en dehors de l'intervalle $[-0.5, 0.5]$. Il y a sûrement des contraintes physiques qui empêchent la variable Y de dépasser certains seuils, et le modèle de régression linéaire n'est pas adapté ici parce qu'il ne peut pas tenir compte de cela.

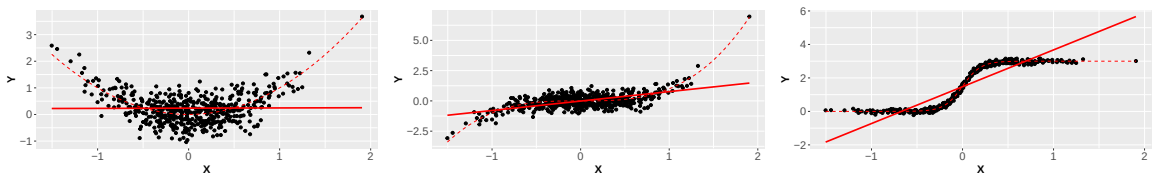


Figure 4.1 – *Application d'un modèle de régression linéaire sur un nuage de points pour lequel la liaison est en réalité non-linéaire.*

Pour étudier la liaison entre deux variables, avec une approche non-linéaire, on introduit dans ce document deux méthodes différentes. La première méthode consiste à effectuer une transformation des données (changement de variable), ce qui permet de pouvoir réutiliser directement les notions introduites dans le chapitre 3. La seconde méthode repose sur l'écriture d'un nouveau modèle, légèrement différent et plus complexe que celle du modèle de régression linéaire simple. Dans ce cas, il est nécessaire d'utiliser en pratique une approche différente pour être capable de déterminer les valeurs des paramètres qui minimisent le critère des moindres carrés.

L'objectif de ce chapitre est de comprendre la différence entre les approches linéaire et non-linéaire, ainsi que d'être capable de mettre en place en pratique un modèle non-linéaire, d'estimer les paramètres et de calculer des

prédictions. Le reste de ce chapitre commence par la section 4.1 dédiée à la méthode par changement de variable. La section 4.2 introduit la méthode par modélisation non-linéaire, dont les détails concernant l'ajustement en pratique sont donnés en section 4.3. Pour finir la section 4.4 contient les diapos de cours ainsi que les feuilles de TD.

Table des matières de ce chapitre

4.1 Changement de variable

L'approche détaillée dans cette section consiste à effectuer un changement de variable sur les données x et/ou y. Pour commencer, donnons une définition de ce qu'est un changement de variable dans ce contexte (contexte différent de celui d'un calcul intégral).

Définition 4.1.1 (Changement de variable) *Pour un vecteur de données $x = (x_1, \dots x_n)$, et une fonction f, effectuer un changement de variable de x avec la fonction f revient à calculer un nouveau vecteur $z = (z_1, \dots, z_n)$ tel que pour chaque $i = 1, \dots, n$,*

$$z_i = f(x_i).$$

Lorsqu'on effectue une changement de variable, en passant d'un vecteur de données x à un vecteur de données z, cela présuppose qu'on pense travailler avec le vecteur z plutôt qu'avec le vecteur x et qu'il y ait plus de chances qu'on obtienne des résultats plus intéressants avec ces nouvelles données. L'objectif derrière cette transformation est qu'on puisse changer l'échelle des données de sorte à ce que la liaison entre les deux variables X et Y puisse être étudiée à l'aide d'un modèle de régression linéaire.

Par exemple, supposons qu'on dispose de données représentées dans le graphique de gauche de la figure 4.2. On constate que la droite de régression ne permet pas de décrire assez fidèlement le nuage de points. Etant donné la forme du nuage de points, on peut penser qu'une transformation logarithmique peut être adaptée, autrement dit un changement de variable avec $f(x) = \log(x)$. Le graphique de droite montre ce qu'on obtient si on étudie la liaison entre $Z(= \log(X))$ et Y. On constate que dans ce cas-là, le modèle de régression linéaire est tout à fait adapté. Avec cette approche, on réussit donc à déterminer indirectement la liaison entre X et Y. En particulier, il s'agit d'une liaison non-linéaire : une liaison "log-linéaire".

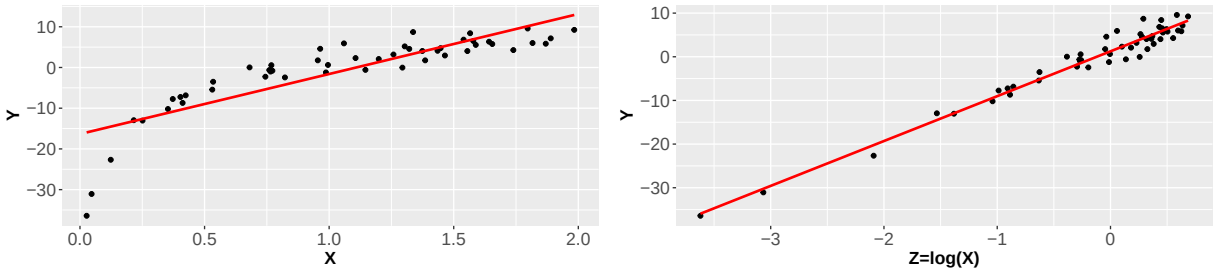


Figure 4.2 – *Application du modèle de régression linéaire sur les données brutes et sur les données transformées.*

Avec un autre exemple, on peut se rendre compte qu'on peut tenir le même raisonnement avec un changement de variable, non pas sur la variable X, mais sur la variable Y. La figure 4.3 donne les résultats d'une approche linéaire sur les données brutes, et sur les données pour lesquelles un changement de variable a été effectué sur la variable Y. Le changement de variable en question ici est $Z = Y^2$.

Le même raisonnement peut être poursuivi jusqu'à appliquer un changement de variable à la fois sur la variable X et à la fois sur la variable Y. Le but d'utiliser un changement de variable sur telle ou telle variable est de faire en sorte d'obtenir un nuage de points qui soit le plus possible "linéarisé", autrement dit qu'il soit relatif à une liaison linéaire. Que l'on choisisse d'effectuer un ou plusieurs changements de variables, il est possible de reformuler un modèle de régression en tenant compte du changement de variable. On donne ci-dessous, un énoncé général du modèle de régression linéaire, tenant compte de potentiel changement de variables.

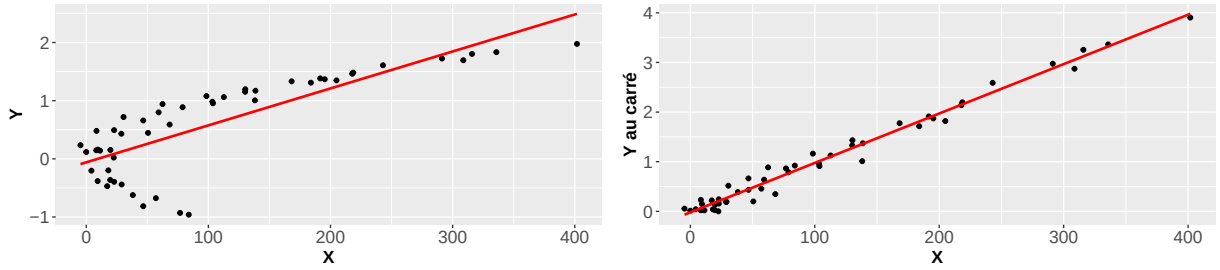


Figure 4.3 – *Application du modèle de régression linéaire sur les données brutes et sur les données transformées.*

Définition 4.1.2 (Modèle de régression linéaire simple avec changement de variable) *Pour des données $x = (x_1, \ldots, x_n)$ et $y = (y_1, \ldots, y_n)$, et de potentiels changements de variable dont les fonctions sont notées f et g, le modèle de régression linéaire simple s'écrit :*

$$g(y_i) = \mu + \beta f(x_i) + \varepsilon_i.$$

Dans le cas de ce modèle, les estimateurs des paramètres μ et β sont donnés par la minimisation du critère des moindres carrés suivant :

$$C(\mu, \beta) = \sum_{i=1}^{n} (g(y_i) - (\mu + \beta f(x_i)))^2$$

et les formules des estimateurs sont :

$$\hat{\mu} = \overline{g(y)} - \hat{\beta}\overline{f(x)} \quad \text{et} \quad \hat{\beta} = \frac{\overline{g(y)f(x)} - \overline{g(y)} \times \overline{f(x)}}{\overline{f(x)^2} - \overline{f(x)}^2}$$

Voici ci-dessous, une liste des fonctions usuelles à utiliser pour trouver un changement de variable adapté, que ce soit pour la fonction f ou la fonction g :

- $f(x) = \exp\{x\}$,
- $f(x) = \log(x)$,
- $f(x) = \sqrt{x}$,
- $f(x) = \frac{1}{x}$,
- $f(x) = x^p$ pour une valeur donnée de p, et
- $f(x) = \sin(x)$ ou $f(x) = \cos(x)$.

En pratique, afin de trouver quel est le bon changement de variable à effectuer (s'il y en a un à faire), il faut soit reconnaître la forme d'une fonction connue dans le nuage de points, soit essayer les différentes fonctions usuelles et de vérifier si on obtient un ajustement linéaire acceptable.

4.2 Modèle non-linéaire

En opposition à l'approche introduite en section 4.1, qui peut s'interpréter comme le fait d'établir une variante du modèle de régression linéaire simple, dans cette section on détaille une approche basée sur un modèle de régression non-linéaire. Ce type de modèle s'écrit de la manière suivante.

Définition 4.2.1 (Modèle de régression non-linéaire) *Pour des données $x = (x_1, \ldots, x_n)$ et $y = (y_1, \ldots, y_n)$, et une fonction f admettant un ou plusieurs paramètres qu'on note par le vecteur θ, le modèle de régression non-linéaire s'écrit comme :*

$$y_i = f(x_i, \theta) + \varepsilon_i.$$

Définition 4.2.2 (Fonction de lien) *Pour le modèle de régression non-linéaire 4.2.1, la fonction f s'appelle la fonction de lien.*

Remarque 4.2.3 (Fonction de lien et régression linéaire)**.** *Pour un modèle de régression linéaire simple, la fonction de lien est la fonction suivante :*

$$f(x_i, \mu, \beta) = \mu + \beta x_i.$$

Ce modèle englobe une grande variété de liaisons possibles entre X et Y et voici ci-dessous les modèles non-linéaires au programme de cette ressource pédagogique :

- le **modèle exponentiel** dont l'équation est :

$$y_i = a + b\exp\{cx_i\} + \varepsilon_i$$

dont les paramètres à estimer sont a, b et c, et on note $\theta = (a, b, c)$.
- le **modèle logistique** dont l'équation est :

$$y_i = \frac{a}{1 + \exp\{\frac{b-x}{c}\}} + \varepsilon_i$$

dont les paramètres à estimer sont a, b et c, et on note $\theta = (a, b, c)$.
- le **modèle polynomial** d'ordre 3 dont l'équation est :

$$y_i = a + bx + cx^2 + dx^3 + \varepsilon_i$$

dont les paramètres à estimer sont a, b, c et d, et on note $\theta = (a, b, c, d)$. Ce modèle peut se décliner suivant le choix de l'ordre choisi pour le polynome. De plus, si le polynome choisi est d'ordre p, il y a $p+1$ paramètres à estimer.

La figure 4.4 donne des illustrations de nuages de points qui sont à juste titre modélisés par chacun de ces modèles. De gauche à droite, les graphiques correspondent aux modèles exponentiel, logistique et polynomial d'ordre 3.

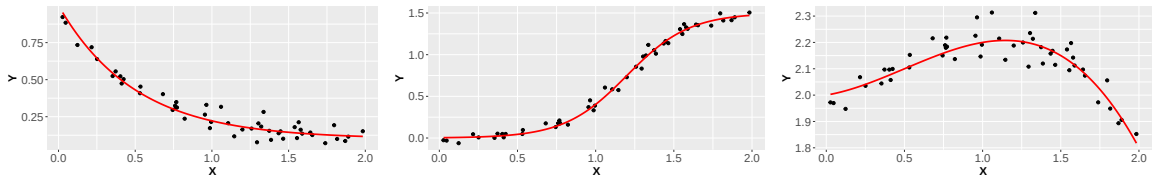


Figure 4.4 – *Exemples de nuages de points pour lesquels une modélisation non-linéaire est adaptée.*

La principale différence entre ces modèles non-linéaires et les variantes du modèle linéaire présentées en section 4.1, est qu'il n'est pas possible d'obtenir une équation de modèle non-linéaire, juste en transformant les données x_i ou y_i dans l'équation du modèle linéaire. Il ne s'agit donc pas d'une variante, mais plutôt d'une généralisation à d'autres formes de liaison. Parmi les éléments qui changent par rapport au modèle linéaire, il y a la notion de courbe de régression.

Définition 4.2.4 (Courbe de régression) *Pour un modèle de régression non-linéaire, la courbe de régression désigne la courbe d'équation :*

$$y = f(x, \hat{\theta})$$

où $\hat{\theta}$ est l'estimation du ou des paramètres θ.

De plus, pour calculer des prédictions on utilise la formule suivant :

$$\hat{y}_i = f(x_i, \hat{\theta}).$$

4.3 Méthode d'ajustement

Bien que ce modèle paraisse pratique puisqu'il permet de modéliser des formes de liaison plus complexes qu'une simple équation de droite, il s'avère qu'appliquer cette modélisation dans un cas concret puisse être compliqué. Une des raisons à cela est que contraitement au modèle de régression linéaire, on ne dispose pas des expressions des estimateurs. Cela vient du fait que dès lors que la forme de la fonction de lien f du modèle non-linéaire est trop complexe, les calculs des dérivées partielles rendent difficile de trouver les points critiques. Pour calculer les estimations à partir des données, il est alors nécessaire d'employer une méthodologie dédiée.

Pour ce modèle de régression non-linéaire, pour estimer les paramètres, il est nécessaire de minimiser le critère des moindres carrés suivant :

$$C(\theta) = \sum_{i=1}^{n} \left(y_i - f(x_i, \theta)\right)^2.$$

Pour trouver la valeur de θ qui minimise $C(\theta)$, on utilise une méthode de descente de gradient (à l'aide d'un ordinateur), et pour commencer on définit ce qu'est un gradient.

Définition 4.3.1 (Gradient) *Pour une fonction à plusieurs variables, $f(x_1, \ldots, x_p)$, le gradient $\nabla f(x)$ de la fonction f en un point $x = (x_1, \ldots, x_p)$ est le vecteur des dérivées partielles :*

$$\nabla f(x) = \begin{pmatrix} \frac{\partial f}{\partial x_1}(x_1, \ldots, x_p) \\ \frac{\partial f}{\partial x_2}(x_1, \ldots, x_p) \\ \vdots \\ \frac{\partial f}{\partial x_p}(x_1, \ldots, x_p) \end{pmatrix}$$

Pour interpréter ce gradient, on peut y voir une similitude avec le coefficient directeur de la tangente à la courbe, autrement dit la valeur de la dérivée. Lorsque la fonction est multidimensionnelle, le gradient donne un "coefficient directeur" du plan tangent à la fonction f. Pour le formuler différement, le gradient peut se comprendre comme : "partant d'un point donné, la direction vers laquelle la fonction f augmente le plus". Pour introduire la méthode de descente de gradient, l'intuition est qu'elle repose sur le fait de chercher la direction de plus forte descente (direction opposée à celle du gradient) et d'entamer une courte descente. L'algorithme recommence alors à rechercher la plus forte descente pour le point auquel on est arrivé avec la descente précédente. En descendant successivement, on doit atteindre finalement un minimum de la fonction.

Définition 4.3.2 (Algorithme de descente de gradient) *A partir d'une valeur initiale x_0, fixée a priori, et un seuil de tolérance $\varepsilon \geq 0$, on exécute les étapes suivantes :*

1. *A partir de la $k^{\text{ème}}$ valeur, calculer le gradient $\nabla f(x_k)$.*
2. *Vérifier la condition d'arrêt : si $\| \nabla f(x_k) \| \leq \varepsilon$, alors on arrête l'algorithme.*
3. *Calculer le pas de descente $\alpha_k > 0$.*
4. *Déterminer la nouvelle valeur : $x_{k+1} = x_k - \alpha_k \nabla f(x_k)$.*

Lorsque la condition d'arrêt provoque l'arrêt de l'algorithme, on récupère la valeur x_k, ce qui correspond à une approximation d'un minimum local de la fonction f.

Pour cette ressource pédagogique, il n'est pas nécessaire de savoir mettre en place par soi-même cet algorithme. Il est seulement requis de savoir utiliser la fonction de R qui cherche à minimiser le critère des moindres carrés dans le cas d'un modèle de régression non-linéaire, et de comprendre les intuitions mathématiques et géométriques derrière cet algorithme.

4.3.1 Mise en pratique

En pratique, pour utiliser avec R cette procédure de minimisation du critère des moindres carrés, il faut utiliser la fonction `nls`. Voici ci-dessous un exemple d'utilisation.

```
# Importer les données de Boston incluse dans la library MASS
library(MASS)

# Tracer une représentation graphique des deux variables d'interet
plot(nox~dis,Boston)

# Ecrire dans R l'équation d'un modèle non-linéaire (exponentiel)
modele_exponentiel <- function(x,a,b,c){
  a+b*exp(c*x)
}

# Ajuster le modèle
res_nls <- nls(nox~modele_exponentiel(dis,a,b,c),data=Boston,
    start = list(a=0.5,b=0.5,c=-1)
)
# Obtenir les valeurs des estimations
theta_estimate <- res_nls$m$getPars()

# Calculer la courbe de régression
x <- seq(min(Boston$dis),max(Boston$dis),le=1e3)
```

```
y <- modele_exponentiel(x,
                        theta_estimate[1],theta_estimate[2],theta_estimate[3]
                        )

# Tracer la courbe de régression
plot(nox~dis,Boston)
lines(x,y,col="red",lwd=2)
```

Pour utiliser la fonction `nls` il est nécessaire de donner les arguments suivants :
- une formule écrite de la forme `y~f(x,theta)` où `f` est une fonction connue de R ou qui a été préalablement définie,
- un objet (data.frame) contenant les données, et
- une liste donnant une valeur numérique pour chacun des paramètres, et qui servent de valeurs initiales pour l'algorithme.

L'exécution de ces lignes de code dans l'exemple ci-dessus ne pose pas de problème mais pourtant, lorsqu'on utilise la fonction `nls` il peut arriver qu'on fasse face à différents problèmes.

4.3.2 Des erreurs possibles

Une des erreurs possibles, et arrivant fréquemment, peut être obtenue avec la commande suivante :

```
res_nls <- nls(nox~modele_exponentiel(dis,a,b,c),data=Boston,
               start = list(a=1,b=1,c=-1)
)
```

ce qui renvoie l'erreur suivante :

```
Error in nls(nox ~ modele_exponentiel(dis, a, b, c), data = Boston, start = list(a = 1,  :
  le pas 0.000488281 est devenu inférieur à 'minFactor' de 0.000976562
```

ce qui indique que le pas α_k qui a été calculé est trop petit. Ceci peut arriver lorsque la fonction de lien f est complexe, et cela dépend aussi de la valeur initiale qu'on a choisi. Pour s'en prévenir, il est convient de changer de valeur initiale, ou alors de définir un niveau minimal plus bas pour la valeur du pas α_k, avec un argument suivant de la fonction `nls` :

```
control = nls.control(minFactor = 0.000001)
```

Cependant, pour la première solution comme la deuxième, il n'y a pas de garantie de trouver les valeurs adaptées de sorte à ce que la fonction `nls` ne renvoie pas d'erreur.

Une deuxième erreur possible, peut-être obtenue avec la commande suivante :

```
res_nls <- nls(nox~modele_exponentiel(dis,a,b,c),data=Boston,
               start = list(a=1,b=-1,c=-1)
)
```

ce qui renvoie l'erreur suivante :

```
Error in nls(nox ~ modele_exponentiel(dis, a, b, c), data = Boston, start = list(a = 1,  :
  gradient singulier
```

ce qui signifie que pour cette valeur de x_k, il n'est pas possible de calculer le gradient du critère des moindres carrés. Tout comme l'erreur précédente, ceci peut arriver lorsque la fonction de lien f est complexe, et cela dépend aussi de la valeur initiale qu'on a choisi. Résoudre cette erreur se fait donc de la même manière.

4.3.3 Valeurs de départ par défaut

Dans certains cas, on peut résoudre ces problèmes en choisissant correctement un point de départ. Cela peut soit se faire par le calcul à la main (hors programme de cette ressource pédagogique), soit pour certains modèles à l'aide d'une fonction de R. La fonction `nls` détermine de manière automatique des valeurs initiales adaptées si on utilise un modèle déjà implémenté dans R. Par exemple, pour utiliser le modèle logistique, il faut utiliser la fonction `SSlogis`. Voici ci-dessous, un code permettant de simuler des données et d'ajuster un modèle logistique sur ces données :

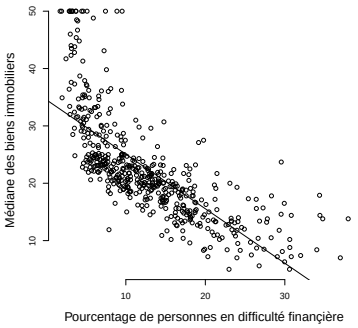
```
# Se fixer une graine pour simuler des données
set.seed(1)

# Simuler des données
x <- runif(50,0,1)
y <- SSlogis(x,1,0.5,0.05) + rnorm(50,0,0.1)
data <- data.frame(x=x,y=y)

# Voir à quoi ressemblent les données
plot(y~x,data=data)

# Ajuster un modèle non-linéaire (logistique)
res_nls <- nls(y~SSlogis(x,a,b,c),data=data)

# Obtenir les valeurs des estimations
theta_estimate <- res_nls$m$getPars()

# Calculer la courbe de régression
x <- seq(min(data$x),max(data$x),le=1e3)
y <- SSlogis(x,
             theta_estimate[1],theta_estimate[2],theta_estimate[3]
)

# Tracer la courbe de régression
plot(y~x,data)
lines(x,y,col="red",lwd=2)
```

4.4 Diapos de cours et exercices de travaux dirigés

Chap. 3 – **Ajustement non-linéaire**

Motivations et objectifs

Contexte :
On dispose de mesures pour deux variables quantitives $x_1, \ldots, x_n$ et $y_1, \ldots, y_n$.
On cherche à déterminer le lien entre ces deux variables.



Pour cela, on a commencé par utiliser le modèle de régression linéaire.
On réalise au regard du nuage de point et du coefficient de détermination que ce modèle n'est pas adapté.

$$R^2 = 0.54.$$

Ce coefficient indique qu'il n'y a pas de liaison linéaire entre ces deux variables.
Il faut chercher à déterminer la liaison qui n'est pas linéaire.
Question :
Comment déterminer un modèle qui ne soit pas une régression **linéaire** ?
Approche :
Pour cela, deux approches sont possibles :

1. Utiliser une régression linéaire sur des modifications des données x et/ou y
2. Modifier le modèle de régression → régression non-linéaire

Détermination, corrélation et linéarité

Détermination et corrélation :
Le coefficient de détermination R^2 est lié à la corrélation entre les deux ensembles de mesures $x = (x_1, \ldots, x_n)$ et $y = (y_1, \ldots, y_n)$:

$$R^2 = r(x, y)^2.$$

Donc, si la corrélation est proche de 0, alors le coefficient de détermination est aussi proche de 0.

Rappels : Corrélation et linéarité
La corrélation (qui est entre -1 et 1) est d'autant plus éloignée de 0 que la forme du nuage de points a une forme linéaire.

Conclusion :
Le coefficient de détermination, tout comme la corrélation, ne capte qu'une linéarité du nuage de points.

Exemples de nuages de points et corrélations

Exemples de liaison non-linéaire

Transformation des données

Principe :
Pour ajuster une variation de la régression linéaire, il est possible de faire une transformation des données. Par exemple, la variable y n'est pas prédite par la variable x, mais par une transformation $f(x)$ de cette variable.

Régression linéaire :
Le modèle est :

$$y_i = \mu + \beta x_i + \varepsilon_i$$

La commande R pour ajuster ce modèle est :

Transformation exponentielle :
On transforme les données x_i : $x \mapsto \exp(x)$

$$y_i = \mu + \beta \exp(x_i) + \varepsilon_i$$

La commande R pour ajuster ce modèle est :

.
.
.
.
.
.
.
.
.
.
.
.
.
.
.
.
.
.
.
.
.
.
.

Autres transformations

Principe :
Pour ajuster une variation de la régression linéaire, il est possible de faire une transformation des données. Par exemple, au lieu de prédire la variable y à par la variable x, on peut prédire une transformation $f(y)$ de la variable y.

Transformation : $y \mapsto \exp(y)$

$$\exp(y_i) = \mu + \beta x_i + \varepsilon_i \iff y_i = \log(\mu + \beta x_i + \varepsilon_i)$$

Transformation : $y \mapsto \log(y)$

$$\log(y_i) = \mu + \beta x_i + \varepsilon_i \iff y_i = \exp(\mu + \beta x_i + \varepsilon_i)$$

Transformation : $y \mapsto y^2$

$$y_i^2 = \mu + \beta x_i + \varepsilon_i \iff y_i = \sqrt{|\mu + \beta x_i + \varepsilon_i|}$$

Transformation : $y \mapsto \sqrt{y}$

$$\sqrt{y_i} = \mu + \beta x_i + \varepsilon_i \iff y_i = (\mu + \beta x_i + \varepsilon_i)^2$$

Transformation : $y \mapsto \frac{1}{y}$

$$\frac{1}{y_i} = \mu + \beta x_i + \varepsilon_i \iff y_i = \frac{1}{\mu + \beta x_i + \varepsilon_i}$$

Pour ajuster ces modèles, il faut utiliser les commandes suivantes :

.
.
.
.
.
.
.
.

Autres transformations

Principe :
Pour ajuster une variation de la régression linéaire, il est possible de faire une transformation des données. Par exemple, au lieu de prédire la variable y à par la variable x, on peut prédire une transformation $f(y)$ de la variable y.

Finalement, d'autres modèles sont possibles en effectuant simultanément des tranformations f_1 et f_2 sur les variables y et x. Le modèle obtenu est :

.
.
.
.
.
.
.
.

Modèles non-linéaires standards

L'approche précédente à consister à transformer les données y et/ou x pour obtenir une variation du modèle de régression linéaire.

Modèle de régression non-linéaire :
Une autre approche possible est d'ajuster un modèle non-linéaire pour un paramètre θ :

$$y_i = f(x_i, \theta) + \varepsilon_i$$

pour une fonction f donnée. Pour ajuster ce type de modèle avec R, on n'utilise pas la fonction `lm` mais une autre fonction plus complexe : `nls`.

Modèle exponentiel :
Le modèle exponentiel à 3 paramètres est donné par :

$$y_i = a + b \times \exp(cx_i) + \varepsilon_i$$

Ce modèle est parfois utilisé sans les paramètres a et d.
Modèle logistique :
Le modèle logistique à 3 paramètres est donné par :

$$y_i = \frac{a}{1 + \exp(\frac{b-x}{c})} + \varepsilon_i$$

Modèles non-linéaires standards

L'approche précédente à consister à transformer les données y et/ou x pour obtenir une variation du modèle de régression linéaire.

Modèle de régression non-linéaire :
Une autre approche possible est d'ajuster un modèle non-linéaire pour un paramètre θ :

$$y_i = f(x_i, \theta) + \varepsilon_i$$

pour une fonction f donnée. Pour ajuster ce type de modèle avec R, on n'utilise pas la fonction `lm` mais une autre fonction plus complexe : `nls`.

Modèle polynomial :
Le modèle polynomial à 4 paramètres est donné par :

$$y_i = a + bx + cx^2 + dx^3 + \varepsilon_i$$

Ajustements

Ajustements :
Afin d'ajuster une variation du modèle linéaire ou un modèle de régression non-linéaire, il est possible d'utiliser la méthode des moindres carrés.

Cas 1 : transformation de la variable x
Pour une transformation f, le modèle est : $y = \mu + \beta f(x) + \varepsilon$.
Les prédictions sont obtenues par : $\hat{y} = \hat{\mu} + \hat{\beta} f(x)$.
Le critère des moindres carrés à minimiser est : $\sum_{i=1}^{n} (y_i - (\hat{\mu} + \hat{\beta} f(x)))^2$.
Pour ajuster le modèle, il faut déterminer les valeurs $\hat{\mu}$ et $\hat{\beta}$, de la même manière que pour le modèle de régression linéaire.

Cas 2 : transformation de la variable y
Pour une transformation f ayant une réciproque f^{-1} : $f(y) = \mu + \beta x + \varepsilon$.
Les prédictions sont obtenues par : $\hat{y} = f^{-1}(\hat{\mu} + \hat{\beta} x)$.
Le critère des moindres carrés à minimiser est : $\sum_{i=1}^{n} (y_i - f^{-1}(\hat{\mu} + \hat{\beta} x))^2$.
Pour ajuster le modèle, il faut déterminer les valeurs $\hat{\mu}$ et $\hat{\beta}$, de la même manière que pour le modèle de régression linéaire.

Ajustements

Ajustements :
Afin d'ajuster une variation du modèle linéaire ou un modèle de régression non-linéaire, il est possible d'utiliser la méthode des moindres carrés.

Cas 3 : modèle de régression non-linéaire
Pour une transformation f qui dépend d'un paramètre θ : $y = f(x, \theta) + \varepsilon$.
Les prédictions sont obtenues par : $\hat{y} = f(x, \hat{\theta})$.
Le critère des moindres carrés à minimiser est : $\sum_{i=1}^{n}(y_i - f(x, \hat{\theta}))^2$.
Pour ajuster le modèle, il faut déterminer les valeurs $\hat{\mu}$ et $\hat{\beta}$. Suivant la transformation f, il n'y a pas forcément de solutions calculables pour $\hat{\mu}$ et $\hat{\beta}$.

Lorsqu'il n'y pas de solution facile à trouver, on utilise un algorithme d'optimisation qui calcule une approximation de la solution.

Choix de modèle

Avec tout les modèles possibles, il y a beaucoup de modèles qui semble adaptés pour décrire le nuage de points..
Comment choisir entre plusieurs modèles qui, pour l'œil humain, semblent valides.

Choisir entre les modèles :
Pour choisir entre les modèles, une première approche possible consiste à comparer l'ajustement. Le critère courament utilisé est le MSE (*Mean Square Error* : moyenne des carrés des erreurs) :

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i) = \frac{1}{n}\sum_{i=1}^{n} e_i.$$

Le meilleur modèle est celui qui a le plus faible MSE.

Exemple :
Pour le modèle 1, on obtient les prédictions $\hat{y}^1 = (\hat{y}_1^1, \dots \hat{y}_n^1)$.
Pour le modèle 2, on obtient les prédictions $\hat{y}^2 = (\hat{y}_1^2, \dots \hat{y}_n^2)$.
Les MSE pour chacun des deux modèles sont :

$$\text{MSE}_1 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i^1)^2 \qquad \text{MSE}_2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i^2)^2$$

Si $\text{MSE}_1 < \text{MSE}_2$ alors on en conclut que le modèle 1 est plus adapté que le modèle 2 pour décrire la liaison entre x et y.
Sinon, l'autre modèle est le plus adapté.

Résumé

Les éléments qu'il faut avoir compris :
- S'extraire de la linéarité lorsque la régression linéaire est inadaptée
  Se voit sur le nuage de points ou avec R^2
- Les liaisons non-linéaires
- Les transformations des données y et/ou x
- Les modèles linéaires
- Comment ajuster ces modèles
  moindres carrés et commandes R pour les transformations
- Choisir un modèle au regard de l'ajustement
  MSE, erreur de prédiction

TD 2 – Régression linéaire et non-linéaire

Les graphiques relatifs à chacun de trois exercices de cette feuille de TD sont en page 4.

Pour la résolution de certaines des questions des exercices, vous aurez besoin de formules et de notations données dans la feuille : ***TD 1 – Optimisation et moindres carrés****.*

Exercice 1. Le but de cet exercice est d'ajuster un nuage de points. Pour ajuster ce nuage de points, on se propose de comparer deux modèles : 1) le modèle de régression linéaire et 2) le même modèle mais avec une transformation exponentielle des données x. Pour la suite, on notera $f(x) = \exp(x)$. Le tableau ci-dessous présente ces données.

	x	y
1	-2.34	0.95
2	-1.28	2.39
3	0.73	3.02
4	4.08	8.91
5	-2.98	1.40
6	3.98	9.28
7	4.45	11.37
8	1.61	2.04
9	1.29	0.28
10	-4.38	2.69

1. Ecrivez l'équation des deux modèles. Les paramètres du premier modèle seront notés μ_1 et β_1 et ceux du second modèle seront notés μ_2 et β_2.
2. Calculez les quantités suivantes à partir des données : $\overline{x}$, $\overline{y}$, $\overline{xy}$, $\overline{x^2}$, $\overline{f(x)}$, $\overline{yf(x)}$ et $\overline{f(x)^2}$.
3. Calculez les estimations des paramètres de chacun des deux modèles.
4. Afin de tracer le graphique relatif au premier modèle, calculez les prédictions du premier modèle pour $x = -2$ et $x = 4$. Reportez les points sur le graphique et tracez la courbe du modèle de régression linéaire.
5. Faites de même pour le second modèle pour $x = -4$, $x = -2$, $x = 0$, $x = 2$ et $x = 4$.
6. Au regard des courbes des deux modèles, quel modèle vous semble être le plus adapté pour ajuster le nuage de points ?
7. Déterminez les prédictions du premier modèle à partir des données x_i, et faites de même pour le second modèle. Celles du premier modèle seront notées $\hat{y}_i^1$ et celles du second modèle $\hat{y}_i^2$.
8. Calculez les résidus de chacun des modèles.
9. En déduire le MSE pour chacun des modèles.
10. Quel modèle s'ajuste le mieux aux données ?

Exercice 2. Pour cet exercice, trois modèles sont en compétition pour ajuster le nuage de points : 1) une régression linéaire, 2) avec une transformation exponentielle des données x et 3) avec une transformation au carré des données x. Pour la suite de cet exercice, on notera $f(x) = \exp(x)$ et $g(x) = x^2$. Pour la résolution de cet exercice vous aurez besoin de certains résumés statistiques comme $\overline{y} = 0.129$ et les autres résumés sont donnés dans le tableau ci-dessous :

Régression linéaire	Transformation $\exp(x)$	Transformation x^2
$\overline{x} = 0.607$ et $\overline{y} = 0.13$	$\overline{f(x)} = 15.045$	$\overline{g(x)} = 6.243$
$\overline{yx} = -2.846$	$\overline{yf(x)} = -60.043$	$\overline{yg(x)} = -12.585$
$\overline{x^2} = 6.243$	$\overline{f(x)^2} = 1097.769$	$\overline{g(x)^2} = 78.756$

1. Donnez l'équation de chacun des modèles. Les paramètres du $j^{\text{ème}}$ modèle seront notés μ_j et β_j.

2. Pour chaque modèle, déterminez les estimations des paramètres à partir des résumés statistiques.
3. Comme il serait trop long de calculer les prédictions pour chacun des modèles et chacune des données x_i, le tableau ci-dessous donne un échantillon de 10 données.

	x	y
1	-0.58	1.67
2	3.89	-1.84
3	0.77	4.81
4	2.02	-0.27
5	-0.06	1.80
6	1.36	1.36
7	-1.85	1.08
8	1.70	1.26
9	2.15	1.97
10	3.78	-3.16

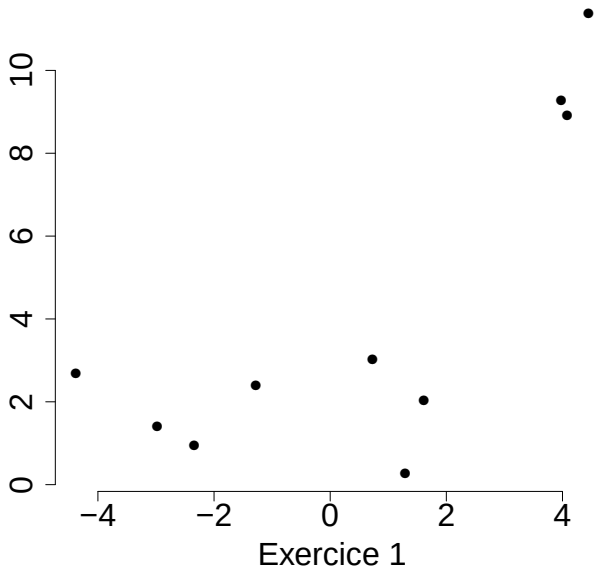
   A partir de ces données x_i, calculez les prédictions $\hat{y}_i$ pour chacun des trois modèles.
4. En déduire les résidus de chacun des modèles pour ces prédictions.
5. Calculez le MSE à partir de ces 10 données pour chacun des modèles.
6. Au regard de ces résultats, quel modèle vous semble le mieux adapté ?
7. Calculez les prédictions du premier modèle pour les données $x = -2$ et $x = 4$. Puis, tracez la courbe relative à ce modèle.
8. Calculez les prédictions du deuxième modèle pour les données $x = -4$, $x = -2$, $x = 0$, $x = 2$ et $x = 4$. Puis, tracez la courbe relative à ce modèle.
9. Calculez les prédictions du troisième modèle pour les données $x = -4$, $x = 0$, et $x = 4$. Puis, tracez la courbe relative à ce modèle.

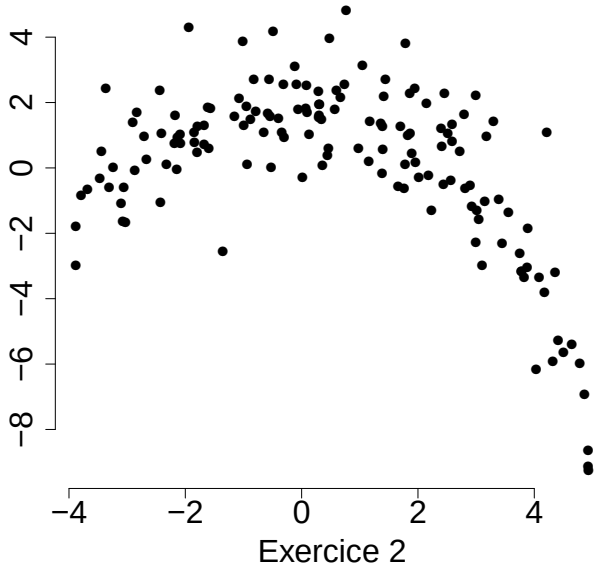
Exercice 3. Dans le cadre de cet exercice, le modèle de régression linéaire est en compétition pour ajuster un nuage de points, avec un modèle de régression non-linéaire : le modèle logistique. Les données à ajuster sont données dans le tableau ci-dessous :

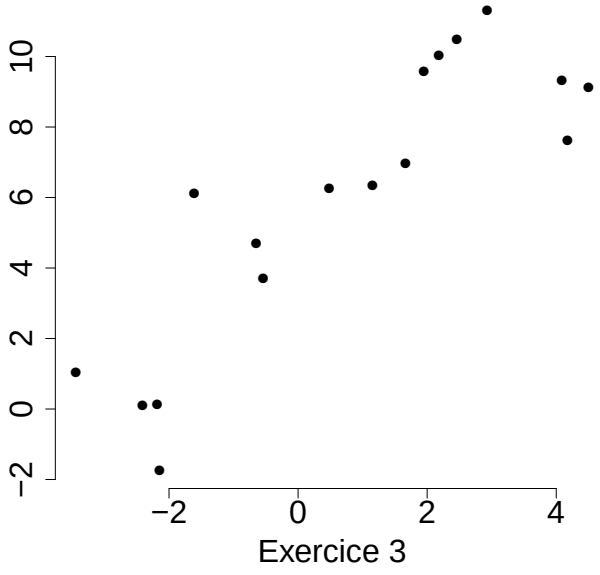
	x	y
1	-1.61	6.12
2	-0.65	4.70
3	1.16	6.36
4	4.17	7.63
5	-2.18	0.13
6	4.09	9.32
7	4.50	9.13
8	1.95	9.59
9	1.66	6.98
10	-3.44	1.04
11	-2.15	-1.73
12	-2.41	0.11
13	2.18	10.04
14	-0.54	3.72
15	2.93	11.30
16	0.48	6.26
17	2.46	10.48

1. Donnez l'équation de chacun des modèles. Pour le modèle de régression linéaire, les paramètres sont notés μ et β. Les trois paramètres du modèle logistique sont notés a, b et c.
2. Déterminez les estimations $\hat{\mu}$ et $\hat{\beta}$ pour le modèle de régression linéaire.
3. Calculez les prédictions de ce modèle pour $x = -2$ et $x = 4$. Tracez ensuite la courbe de ce modèle.

4. Déterminez les prédictions pour les données x_i.
5. Déduisez-en les résidus de ce modèle.
6. Déduisez-en le MSE pour ce modèle.
7. Pour la suite, on suppose que les estimations des paramètres du modèle logistique sont : $\hat{a} = 10.10$, $\hat{b} = 0.12$ et $\hat{c} = 0.79$. Calculez alors les prédictions du modèle pour $x = -3$, $x = -2$, $x = -1$, $x = 0$, $x = 1$, $x = 2$ et$x = 4$. Tracez ensuite la courbe de ce modèle.
8. Calculez les prédiction du modèle logistique pour les données x_i.
9. Déduisez-en les résidus du modèle logistique.
10. Déduisez-en le MSE du modèle logistique.
11. Déterminez lequel des deux modèles est le plus adapté pour ajuster ce nuage de points.


0
2
4
6
8
10
−4
−2
0
2
4
Exercice 1


−8
−6
−4
−2
0
2
4
−4
−2
0
2
4
Exercice 2


−2
0
2
4
6
8
10
−2
0
2
4
Exercice 3

CHAPITRE 5

Introduction au choix de modèles

Les chapitres 3 et 4 ont été l'occasion d'introduire différents modèles pour étudier des données. Pour le modèle de régression linéaire, la notion de qualité d'ajustement est introduite, mais cela ne permet pas de savoir s'il est plus judicieux d'employer ce modèle ou plutôt un modèle de régression non-linéaire. De plus, si un modèle de régression non-linéaire était préférable, lequel parmi les modèles possibles serait à utiliser ? Pour répondre à ces questions, il faut être capable de comparer les modèles entre eux. Autrement dit, il est nécessaire de hiérarchiser les différents modèles possibles en terme de qualité d'ajustement aux données.

Table des matières de ce chapitre

5.1 Critère de comparaison

Pour comparer différents modèles, il existe plusieurs critères, qui sont à utiliser suivant le type de modèles qui sont comparés. Pour les modèle de régression linéaire et non-linéaire, un critère simple à utiliser est le MSE (pour *Mean Square Error* : la moyenne des carrés des erreurs).

Définition 5.1.1 (MSE) *Le* MSE *est donné par la formule suivante :*

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n} e_i^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2.$$

Pour l'interpréter, on peut noter que le MSE quantifie l'erreur de prédiction qui est commis sur l'intégralité des données. De plus, son expression est très similaire au critère des moindres carrés, et en particulier le MSE correspond à la valeur minimale de ce critère. Un autre lien interprétable est la proposition suivante.

Proposition 5.1.2 (Lien entre MSE **et** R^2**)** *En reconnaissant que :*

$$\text{SCR} = n\,\text{MSE}$$

et en utilisant la décomposition de la variance, on peut établir que le MSE *évolue de manière contraire au coefficient de détermination* R^2 *:*

$$R^2 = 1 - \frac{n\,\text{MSE}}{\text{SCT}}$$

Ces différentes informations permettent d'interpréter que

- plus le MSE est faible, plus le modèle permet de prédire correctement les données, et
- plus le MSE est élevé, plus le modèle commet des erreurs pour prédire les données.

5.2 Choix de modèle

Lorsqu'on étudie des données et en particulier la liaison qu'il y a entre deux variables des données, il est souvent difficile de savoir quel modèle utiliser. On note $M_1, M_2, \ldots, M_K$ les K modèles en compétitions pour étudier des

données. Pour chacun de ces modèles on calcule la valeur du critère MSE. La règle de décision qu'on peut se donner est de choisir le modèle M_k pour lequel le MSE est le plus faible. L'intuition est que celui ayant le plus faible MSE, c'est-à-dire celui faisant le moins d'erreurs de prédiction, est celui pour lequel la modélisation doit être la plus proche de la réelle liaison entre les deux variables.

Définition 5.2.1 (Choix de modèle avec MSE) *Pour choisir un modèle parmi une collection de modèles, on garde celui admettant la valeur de* MSE *la plus faible.*

Dès lors qu'un modèle est sélectionné, il est possible de réaliser l'analyse descriptive (prédiction, coefficient de détermination, interprétations), mais il faut garder en tête qu'on ne dispose pas du vrai modèle pour étudier les données, bien qu'on ait sélectionné le meilleur des modèles. Il y a plusieurs explications pour comprendre pourquoi on ne dispose pas forcément d'un bon modèle dans ce cas-là. Pour détailler ces explications, on commence par donner un sens à ce qu'on veut entendre par "vrai modèle" des données.

Définition 5.2.2 (Vrai modèle) *Le vrai modèle relatif aux données, qu'on note M_0, correspond à la fonction de densité conditionnelle $f_{Y|X=x_0}$ (voir la définition 2.2.6). A savoir, connaissant la valeur de la variable $X = x_0$ pour un individu, on connait l'aléatoire de la donnée y_0 pour ce même individu.*

Avoir le vrai modèle des données ne signifie pas qu'on soit capable de prédire la valeur exacte de chaque donnée observé, ainsi que de potentielles données futures. Cela signifie juste qu'on dispose de la caractérisation de l'aléatoire de la "mécanique de production de la donnée". Cet "aléatoire" inclut des aspects imprévisibles de la variable Y :

- non-connaisance d'une potentielle autre variable Z qui serait la cause de la variable Y, ou
- des variations imprévisibles dues à l'imprécision du capteur de mesure.

Remarque 5.2.3 (Prédictions trop proches des données observées)**.** *Il est à noter que si un modèle prédit "trop bien" les données, cela peut aussi poser un problème. Ce problème s'appelle le sur-apprentissage et correspond au fait d'utiliser un modèle trop complexe qui permette de prédire quasi-exactement les valeurs observées. Cela induit d'avoir un modèle qui modélise la part chaotique et imprévisible du à l'échantillonnage et au capteur de mesure, ce qui ne correspond pas vraiment à une modélisation du phénomène d'intérêt. Pour le formuler autrement, on peut imaginer que la "mécanique de production des données" se fait en trois étapes :* 1) *le phénomène d'intérêt provoque un événement,* 2) *l'expérimentateur fait le choix de mesurer cet événement et* 3) *le capteur de mesure quantifie l'événement avec une imprécision qui lui est propre et qui est imprévisible. Si on modélise les étapes* 2) *et* 3)*, on obtient une approche qui englobe des aspects qui ne nous intéresse pas. En étudiant les données, on souhaiterait faire la part des choses entre la variabilité des données due au phénomène d'intérêt, et la variabilité due à l'expérience (échantillonnage et mesure). Or une modélisation faisant face à du sur-apprentissage, prédit avec une trop grande précision les données et ne fait pas la part entre ces différences sources de variabilité.*
Cette notion de sur-apprentissage est détaillée dans un module de deuxième année, il ne s'agit ici que de l'introduction d'une intuition de cette notion.

Exemple 5.2.4 (Prédictions trop proches des données observées)**.** *Supposons disposer de données de températures (y_i) pour des moments d'une même journée entre 12h et 17h (x_i). Les données sont présentées avec le graphique de la figure 5.1. Sur ce graphique, il est aussi représenté avec la courbe noire, une modélisation non-linéaire qui permet de prédire tout les points observés avec exactitude. Il s'agit ici typiquement d'une situation de sur-apprentissage qui ne correspond pas à une bonne modélisation du phénomène. Pour se convaincre de cela, il faut évaluer si des prédictions pour de nouvelles valeurs de x_i donnent des résultats cohérents. Dans ce cas-là, on peut se convaincre du contraire puisque, par exemple, il ne serait pas cohérent d'avoir à 16h un pic de température à 25 degrés alors qu'il fait environ 13 degrés vers 15h30 et environ 14 degrés vers 17h. Autrement dit, garantir de pouvoir prédire quasi-exactement les valeurs observées y_i se fait au prix d'avoir des prédictions incohérentes pour d'autres valeurs de x, ce qui n'est pas souhaitable.*

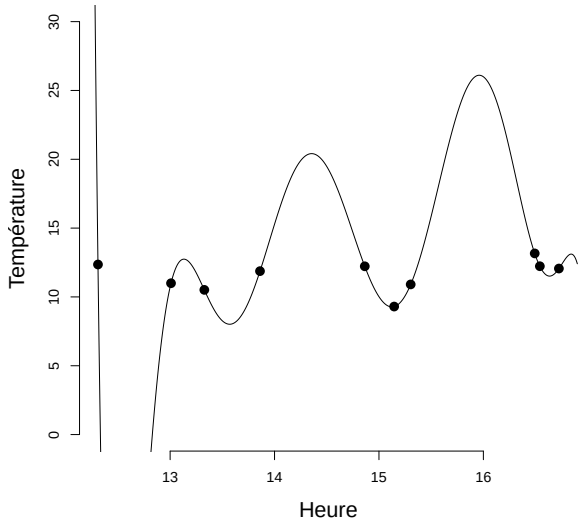


Figure 5.1 – *Données de température sur une demie-journée. La courbe noire correspond à un modèle mathématique permettant de prédire avec exactitude chacune des données observées. Il s'agit d'un exemple de sur-apprentissage.*

Lorsqu'on choisit un modèle parmi les modèles en compétition, on n'est pas assuré que le vrai modèle fasse partie des modèles considérés. Pire encore, on n'est pas assuré qu'il n'y ait ne serait-ce qu'un seul bon modèle dans les modèles considérés. Si c'est le cas, on se sera contenté de choisir le meilleur des mauvais modèles, ce qui n'en fait pas un bon modèle en soi. De plus, admettons que le vrai modèle fasse partie des modèles considérés, il n'arrive pas à coup sûr que cela soit celui-là qui soit sélectionné. Si on ne dispose que de peu de données, on pourrait se retrouver à sélectionner un autre modèle parce qu'il s'ajuste mieux aux données. Que le modèle fasse ou pas partie des modèles considérés, on ne peut donc pas avoir une confiance absolue dans le modèle sélectionné.

De plus, il est à noter qu'en pratique les données (et plus précisément la mécanique de production des données) n'obéissent pas réellement à un modèle mathématique. Les approches mathématique sont des perceptions simplifiées de la réalité, nous permettant de rendre intelligible la complexité de cette réalité. Cela peut se résumer par la citation suivante à garder en tête : "Aucun modèle n'est correct, mais certains son utiles". Etudier des modèles pour analyser des données à un intérêt, et en particulier le fait de sélectionner le meilleur modèle est une étape importante pour s'assurer de fournir des conclusions pertinentes concernant le phénomène d'intérêt. Dans d'autres modules, des critères plus complexes sont introduits afin de réaliser une sélection de modèle de sorte à ce prévenir le plus possible des problèmes évoqués dans ce chapitre.