Chapitre 1

La régression linéaire simple

1.1 Introduction

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

1.1.1 Un exemple : la pollution de l'air

La pollution de l'air constitue actuellement une des préoccupations majeures de santé publique. De nombreuses études épidémiologiques ont permis de mettre en évidence l'influence sur la santé de certains composés chimiques comme le dioxyde

P.-A. Cornillon et al., *Régression avec R*

de souffre (SO_2), le dioxyde d'azote (NO_2), l'ozone (O_3) ou des particules sous forme de poussières contenues dans l'air. L'influence de cette pollution est notable sur les personnes sensibles (nouveau-nés, asthmatiques, personnes âgées). La prévision des pics de concentration de ces composés est donc importante. Nous nous intéressons plus particulièrement à la concentration en ozone. Nous possédons quelques connaissances *a priori* sur la manière dont se forme l'ozone, grâce aux lois régissant les équilibres chimiques. La concentration de l'ozone est fonction de la température ; plus la température est élevée, plus la concentration en ozone est importante. Cette relation très vague doit être améliorée afin de pouvoir prédire les pics d'ozone.
Afin de mieux comprendre ce phénomène, l'association Air Breizh (surveillance de la qualité de l'air en Bretagne) mesure depuis 1994 la concentration en O_3 (en μg/ml) toutes les 10 minutes et obtient donc le maximum journalier de la concentration en O_3, noté dorénavant `O3`. Air Breizh collecte également à certaines heures de la journée des données météorologiques comme la température, la nébulosité, le vent... Les données sont disponibles en ligne (voir Avant-propos). Le tableau suivant donne les 5 premières mesures effectuées.

Individu	O3	T12
1	63.6	13.4
2	89.6	15
3	79	7.9
4	81.2	13.1
5	88	14.1

Tableau 1.1 – 5 données de température à 12 h et teneur maximale en ozone.

Nous allons donc chercher à expliquer le maximum de `O3` de la journée par la température à 12 h. Le but de cette régression est double :

- ajuster un modèle pour expliquer la concentration en `O3` en fonction de `T12` ;
- prédire les valeurs de concentration en `O3` pour de nouvelles valeurs de `T12`.

Avant toute analyse, il est intéressant de représenter les données.

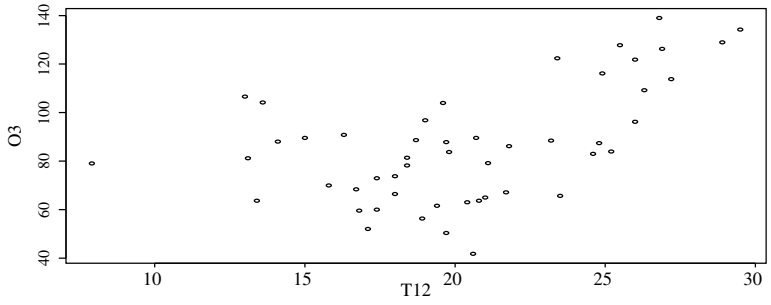

Fig. 1.1 – 50 données journalières de température et `O3`.

Chaque point du graphique (fig.1.1) représente, pour un jour donné, une mesure de la température à 12 h et le pic d'ozone de la journée.
Pour analyser la relation entre les x_i (température) et les y_i (ozone), nous allons chercher une fonction f telle que

$$y_i \approx f(x_i).$$

Pour définir $\approx$, il faut donner un critère quantifiant la qualité de l'ajustement de la fonction f aux données et une classe de fonctions $\mathcal{G}$ dans laquelle est supposée se trouver la vraie fonction inconnue. Le problème mathématique peut s'écrire de la façon suivante :

$$\underset{f \in \mathcal{G}}{\operatorname{argmin}} \sum_{i=1}^{n} l(y_i - f(x_i)), \tag{1.1}$$

où n représente le nombre de données à analyser et $l(.)$ est appelée *fonction de coût* ou encore *fonction de perte*.

1.1.2 Un deuxième exemple : la hauteur des arbres

Cet exemple utilise des données fournies par l'UR2PI et le CIRAD forêt (voir Remerciements). Lorsque le forestier évalue la vigueur d'une forêt, il considère souvent la hauteur des arbres qui la compose. Plus les arbres sont hauts, plus la forêt ou la plantation produit. Si l'on cherche à quantifier la production par le volume de bois, il est nécessaire d'avoir la hauteur de l'arbre pour calculer le volume de bois grâce à une formule du type « tronc de cône ». Cependant, mesurer la hauteur d'un arbre d'une vingtaine de mètres n'est pas aisé et demande un dendromètre. Ce type d'appareil mesure un angle entre le sol et le sommet de l'arbre. Il nécessite donc une vision claire de la cime de l'arbre et un recul assez grand afin d'avoir une mesure précise de l'angle et donc de la hauteur.
Dans certains cas, il est impossible de mesurer la hauteur, car ces deux conditions ne sont pas réunies, ou la mesure demande quelquefois trop de temps ou encore le forestier n'a pas de dendromètre. Il est alors nécessaire d'estimer la hauteur grâce à une mesure simple, la mesure de la circonférence à 1 mètre 30 du sol.
Nous possédons des mesures sur des eucalyptus dans une parcelle plantée et nous souhaitons à partir de ces mesures élaborer un modèle de prévision de la hauteur. Les eucalyptus étant plantés pour servir de matière première dans la pâte à papier, ils sont vendus au volume de bois. Il est donc important de connaître le volume et par là même la hauteur, afin d'évaluer la réserve en matière première dans la plantation (ou volume sur pied total). Les surfaces plantées sont énormes, il n'est pas question de prendre trop de temps pour la mesure et prévoir la hauteur par la circonférence est une méthode permettant la prévision du volume sur pied. La parcelle d'intérêt est constituée d'eucalyptus de 6 ans, âge de « maturité » des eucalyptus, c'est-à-dire l'âge en fin de rotation avant la coupe. Dans cette parcelle, nous avons alors mesuré $n = 1429$ couples circonférence-hauteur. Le tableau suivant donne les 5 premières mesures effectuées.

Individu	ht	circ
1	18.25	36
2	19.75	42
3	16.50	33
4	18.25	39
5	19.50	43

Tableau 1.2 – Hauteur et circonférence (`ht` et `circ`) des 5 premiers eucalyptus.

Nous souhaitons donc expliquer la hauteur par la circonférence. Avant toute modélisation, nous représentons les données. Chaque point du graphique 1.2 représente une mesure du couple circonférence/hauteur sur un eucalyptus.

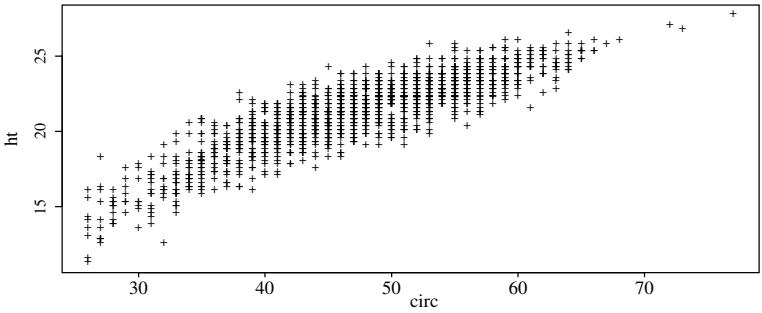

Fig. 1.2 – Représentation des mesures pour les $n = 1429$ eucalyptus mesurés.

Pour prévoir la hauteur en fonction de la circonférence, nous allons donc chercher une fonction f telle que

$$y_i \approx f(x_i)$$

pour chaque mesure $i \in \{1, \ldots, 1429\}$.
A nouveau, afin de quantifier le symbole $\approx$, nous allons choisir une classe de fonctions $\mathcal{G}$. Cette classe représente tous les fonctions d'ajustement possible pour modéliser la hauteur en fonction de la circonférence. Puis nous cherchons la fonction de $\mathcal{G}$ qui soit la plus proche possible des données selon une fonction de coût. Cela s'écrit

$$\underset{f \in \mathcal{G}}{\operatorname{argmin}} \sum_{i=1}^{n} l(y_i - f(x_i)),$$

où n représente le nombre de données à analyser et $l(.)$ est appelée *fonction de coût* ou encore *fonction de perte*.

Remarque
Le calcul du volume proposé ici est donc fait en deux étapes : dans la première on estime la hauteur et dans la seconde on utilise une formule de type « tronc de cône » pour calculer le volume avec la hauteur estimée et la circonférence. Une

autre méthode de calcul de volume consiste à ne pas utiliser de formule incluant la hauteur et prévoir directement le volume en une seule étape. Pour cela il faut calibrer le volume en fonction de la circonférence et il faut donc la mesure de nombreux volumes en fonction de circonférences, ce qui est très coûteux et difficile à réactualiser.

1.2 Modélisation mathématique

Nous venons de voir que le problème mathématique peut s'écrire de la façon suivante (voir équation 1.1) :

$$\underset{f\in\mathcal{G}}{\operatorname{argmin}} \sum_{i=1}^{n} l(y_i - f(x_i)),$$

où $l(.)$ est appelée *fonction de coût* et $\mathcal{G}$ un ensemble de fonctions données. Dans la suite de cette section, nous allons discuter du choix de la fonction de coût et de l'ensemble $\mathcal{G}$. Nous présenterons des graphiques illustratifs bâtis à partir de 10 données fictives de température et de concentration en ozone.

1.2.1 Choix du critère de qualité et distance à la droite

De nombreuses fonctions de coût $l(.)$ existent, mais les deux principales utilisées sont les suivantes :
– $l(u) = u^2$ coût quadratique ;
– $l(u) = |u|$ coût absolu.
Ces deux fonctions sont représentées sur le graphique 1.3 :

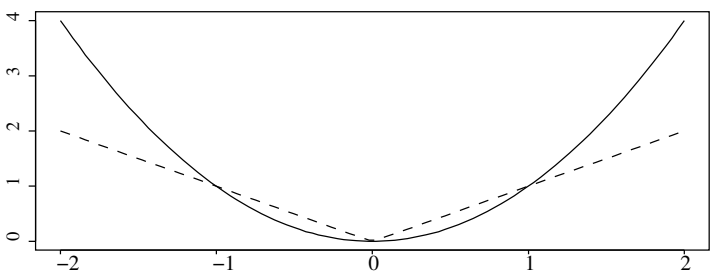

Fig. 1.3 – Coût absolu (pointillés) et coût quadratique (trait plein).

Ces fonctions sont positives, symétriques, elles donnent donc la même valeur lorsque l'erreur est positive ou négative et s'annulent lorsque u vaut zéro.
La fonction l peut aussi être vue comme la distance entre une observation (x_i, y_i) et son point correspondant sur la droite $(x_i, f(x_i))$ (voir fig. 1.4).

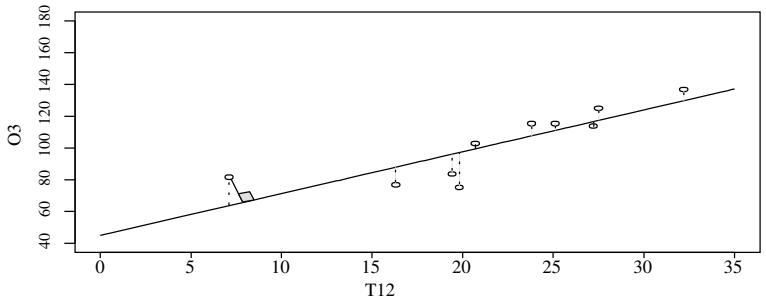

Fig. 1.4 – Distances à la droite : coût absolu (pointillés) et distance d'un point à une droite.

Par point correspondant, nous entendons « évalué » à la même valeur x_i. Nous aurions pu prendre comme critère à minimiser la somme des distances des points (x_i, y_i) à la droite [1] (voir fig. 1.4), mais ce type de distance n'entre pas dans le cadre des fonctions de coût puisqu'au point (x_i, y_i) correspond sur la droite un point $(x_i', f(x_i'))$ d'abscisse et d'ordonnée différentes.
Il est évident que, par rapport au coût absolu, le coût quadratique accorde une importance plus grande aux points qui restent éloignés de la droite ajustée, la distance étant élevée au carré (voir fig. 1.3). Sur l'exemple fictif, dans la classe $\mathcal{G}$ des fonctions linéaires, nous allons minimiser $\sum_{i=1}^{n}(y_i - f(x_i))^2$ (coût quadratique) et $\sum_{i=1}^{n}|y_i - f(x_i)|$ (coût absolu). Les droites ajustées sont représentées sur le graphique ci-dessous :

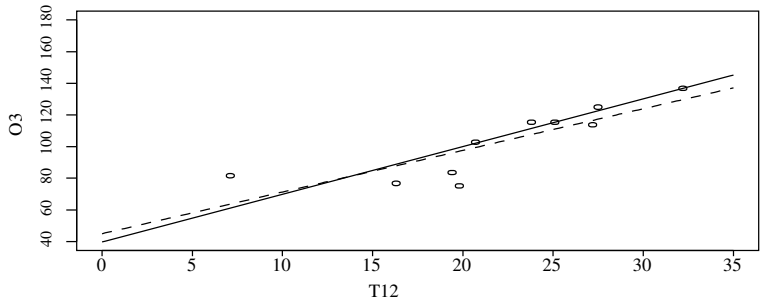

Fig. 1.5 – 10 données fictives de température et O3, régressions avec un coût absolu (trait plein) et quadratique (pointillé).

La droite ajustée avec un coût quadratique propose un compromis où aucun point n'est très éloigné de la droite : le coût quadratique est sensible aux points aberrants qui sont éloignés de la droite. Ainsi (fig. 1.5) le premier point d'abscisse approximative 7°C est assez éloigné des autres. La droite ajustée avec un coût quadratique lui accorde une plus grosse importance que l'autre droite et passe relativement donc plus près de lui. En enlevant ce point (de manière imaginaire),

[1] La distance d'un point à une droite est la longueur de la perpendiculaire à cette droite passant par ce point.

la droite ajustée risque d'être très différente : le point est dit influent et le coût quadratique peu robuste. Le coût absolu est plus robuste et la modification d'une observation modifie moins la droite ajustée. Les notions de points influents, points aberrants, seront approfondies au chapitre 4.
Malgré cette non-robustesse, le coût quadratique est le coût le plus souvent utilisé, cela pour plusieurs raisons : historique, calculabilité, propriétés mathématiques. En 1800, il n'existait pas d'ordinateur et l'utilisation du coût quadratique permettait de calculer explicitement les estimateurs à partir des données. A propos de l'utilisation d'autres fonctions de coût, voici ce que disait Gauss (1809) : « Mais de tous ces principes, celui des moindres carrés est le plus simple : avec les autres, nous serions conduits aux calculs les plus complexes ». En conclusion, *seul le coût quadratique sera automatiquement utilisé dans la suite de ce livre, sauf mention contraire*. Les lecteurs intéressés par le coût absolu peuvent consulter le livre de Dodge & Rousson (2004).

1.2.2 Choix des fonctions à utiliser

Si la classe $\mathcal{G}$ est trop large, par exemple la classe des fonctions continues ($\mathcal{C}_0$), un grand nombre de fonctions de cette classe minimisent le critère (1.1). Ainsi toutes les fonctions de la classe qui passent par tous les points (interpolation), quand c'est possible, annulent la quantité $\sum_{i=1}^{n} l(y_i - f(x_i))$.

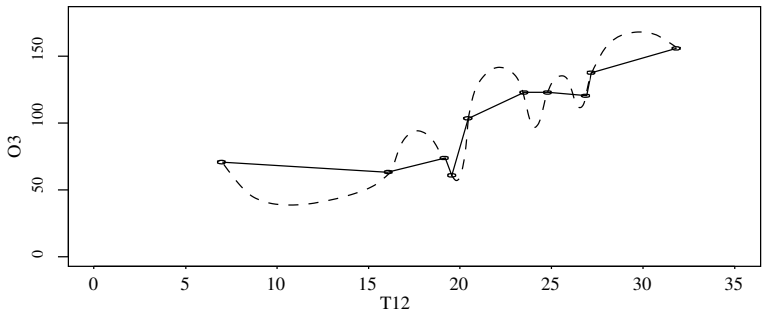

Fig. 1.6 – Deux fonctions continues annulant le critère (1.1).

La fonction continue tracée en pointillés sur la figure (fig. 1.6) semble inappropriée bien qu'elle annule le critère (1.1). La fonction continue tracée en traits pleins annule aussi le critère (1.1). D'autres fonctions continues annulent ce critère, la classe des fonctions continues est trop vaste. Ces fonctions passent par tous les points et c'est là leur principal défaut. Nous souhaitons plutôt une courbe, ne passant pas par tous les points, mais possédant un trajet harmonieux, sans trop de détours. Bien sûr le trajet sans aucun détour est la ligne droite et la classe $\mathcal{G}$ la plus simple sera l'ensemble des fonctions affines. Par abus de langage, on emploie le terme de fonctions linéaires. D'autres classes de fonctions peuvent être choisies et ce choix est en général dicté par une connaissance *a priori* du phénomène et (ou) par l'observation des données.

Ainsi une étude de régression linéaire simple débute toujours par un tracé des observations (x, y). Cette première représentation permet de savoir si le modèle linéaire est pertinent. Le graphique suivant représente trois nuages de points différents.

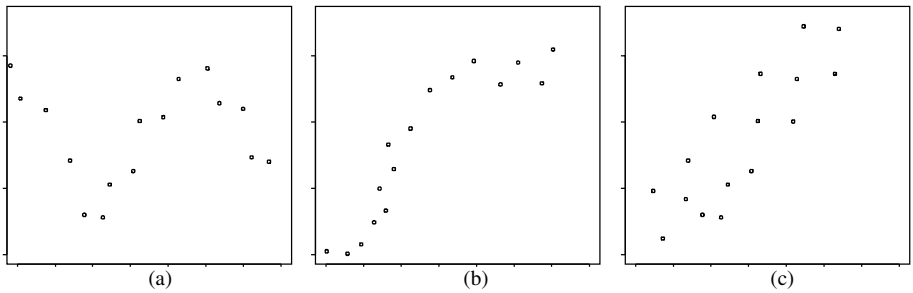

Fig. 1.7 – Exemples fictifs de tracés : (a) fonction sinusoïdale, (b) fonction croissante sigmoïdale et (c) droite.

Au vu du graphique, il semble inadéquat de proposer une régression linéaire pour les deux premiers graphiques, le tracé présentant une forme sinusoïdale ou sigmoïdale. Par contre, la modélisation par une droite de la relation entre X et Y pour le dernier graphique semble correspondre à la réalité de la liaison. Dans la suite de ce chapitre, nous prendrons $\mathcal{G} = \left\{ f : f(x) = ax + b, \quad (a, b) \in \mathbb{R}^2 \right\}$.

1.3 Modélisation statistique

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

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

Dans l'exemple de l'ozone, nous supposons donc un modèle où la concentration d'ozone dépend linéairement de la température. Nous savons pertinemment que toutes les observations mesurées ne sont pas sur la droite. D'une part, il est irréaliste de croire que la concentration de l'ozone dépend linéairement de la température et de la température seulement. D'autre part, les mesures effectuées dépendent de la précision de l'appareil de mesure, de l'opérateur et il peut arriver que pour des valeurs identiques de la variable X, nous observions des valeurs différentes pour Y.
Nous supposons alors que la concentration d'ozone dépend linéairement de la température mais cette liaison est perturbée par un « bruit ». Nous supposons en fait que les données suivent le modèle suivant :

$$Y = \beta_1 + \beta_2 X + \varepsilon. \tag{1.2}$$

L'équation (1.2) est appelée **modèle de régression linéaire** et dans ce cas précis **modèle de régression linéaire simple**. Les β_j, appelés les paramètres du modèle (constante de régression et coefficient de régression), sont fixes mais inconnus, et nous voulons les estimer. La quantité notée ε est appelée bruit, ou erreur, et est aléatoire et inconnue.
Afin d'estimer les paramètres inconnus du modèle, nous mesurons dans le cadre de la régression simple une seule variable explicative ou variable exogène X et une variable à expliquer ou variable endogène Y. La variable X est souvent considérée comme non aléatoire au contraire de Y. Nous mesurons alors n observations de la variable X, notées x_i, où i varie de 1 à n, et n valeurs de la variable à expliquer Y notées y_i.
Nous supposons que nous avons collecté n couples de données (x_i, y_i) où y_i est la réalisation de la variable aléatoire Y_i. Par abus de notation, nous confondrons la variable aléatoire Y_i et sa réalisation, l'observation y_i. Avec la notation ε_i, nous confondrons la variable aléatoire avec sa réalisation. Suivant le modèle (1.2), nous pouvons écrire

$$y_i = \beta_1 + \beta_2 x_i + \varepsilon_i, \qquad i = 1, \cdots, n$$

où
– les x_i sont des valeurs connues non aléatoires ;
– les paramètres β_j, $j = 1, 2$ du modèle sont inconnus ;
– les ε_i sont les réalisations inconnues d'une variable aléatoire ;
– les y_i sont les observations d'une variable aléatoire.

1.4 Estimateurs des moindres carrés

Définition 1.1 (estimateurs des MC)
On appelle estimateurs des moindres carrés (MC) de β_1 et β_2, les estimateurs $\hat{\beta}_1$ et $\hat{\beta}_2$ obtenus par minimisation de la quantité

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

où $\mathbb{1}$ est le vecteur de $\mathbb{R}^n$ dont tous les coefficients valent 1. Les estimateurs peuvent également s'écrire sous la forme suivante :

$$(\hat{\beta}_1, \hat{\beta}_2) = \underset{(\beta_1, \beta_2) \in \mathbb{R} \times \mathbb{R}}{\operatorname{argmin}} \; S(\beta_1, \beta_2).$$

1.4.1 Calcul des estimateurs de β_j, quelques propriétés

La fonction $S(\beta_1, \beta_2)$ est strictement convexe. Si elle admet un point singulier, celui-ci correspond à l'unique minimum. Annulons les dérivées partielles, nous

obtenons un système d'équations appelées équations normales :

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

La première équation donne

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

et nous avons un estimateur de l'ordonnée à l'origine

$$\hat{\beta}_1 = \bar{y} - \hat{\beta}_2 \bar{x}, \tag{1.3}$$

où $\bar{x} = \sum_{i=1}^{n} x_i/n$. La seconde équation donne

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

En remplaçant $\hat{\beta}_1$ par son expression (1.3) nous avons une première écriture de

$$\hat{\beta}_2 = \frac{\sum x_i y_i - \sum x_i \bar{y}}{\sum x_i^2 - \sum x_i \bar{x}},$$

et en utilisant astucieusement la nullité de la somme $\sum(x_i - \bar{x})$, nous avons d'autres écritures pour l'estimateur de la pente de la droite

$$\hat{\beta}_2 = \frac{\sum x_i(y_i - \bar{y})}{\sum x_i(x_i - \bar{x})} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})(x_i - \bar{x})} = \frac{\sum (x_i - \bar{x}) y_i}{\sum (x_i - \bar{x})^2}. \tag{1.4}$$

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

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

Si nous évaluons la droite aux points x_i ayant servi à estimer les paramètres, nous obtenons des $\hat{y}_i$ et ces valeurs sont appelées les valeurs ajustées. Si nous évaluons la droite en d'autres points, les valeurs obtenues seront appelées les valeurs prévues ou prévisions. Représentons les points initiaux et la droite de régression estimée. La droite de régression passe par le centre de gravité du nuage de points $(\bar{x}, \bar{y})$ comme l'indique l'équation (1.3).

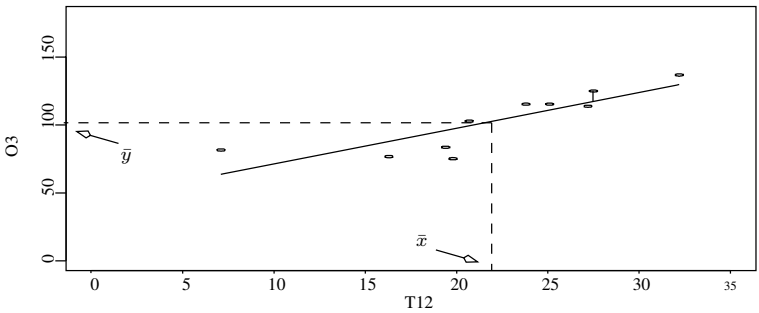

Fig. 1.8 – Nuage de points, droite de régression et centre de gravité.

Nous avons réalisé une expérience et avons mesuré n valeurs (x_i, y_i). A partir de ces n valeurs, nous avons obtenu un estimateur de β_1 et de β_2. Si nous refaisions une expérience, nous mesurerions n nouveaux couples de données (x_i, y_i). A partir de ces données, nous aurions un nouvel estimateur de β_1 et de β_2. Les estimateurs sont fonction des données mesurées et changent donc avec les observations collectées (fig. 1.9). Les vraies valeurs de β_1 et β_2 sont inconnues et ne changent pas.

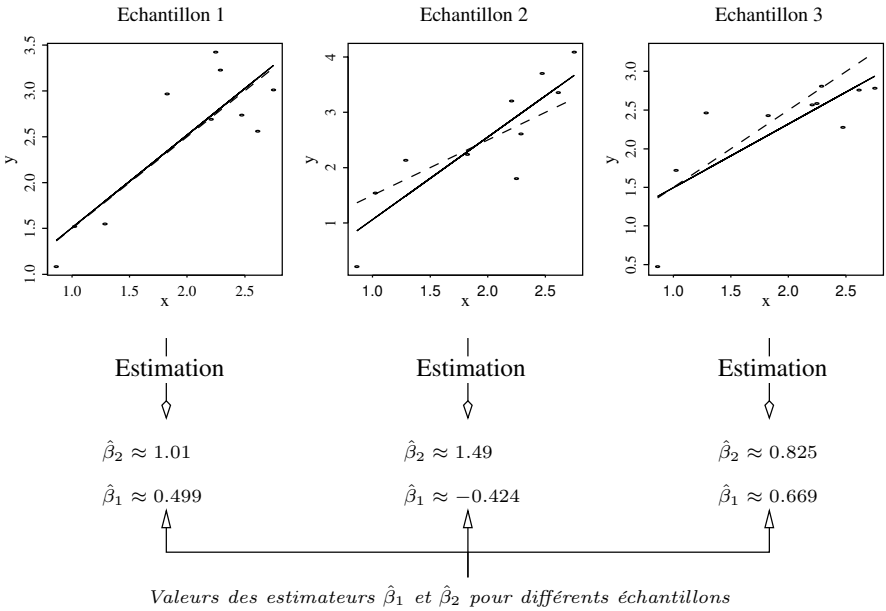

Fig. 1.9 – Exemple de la variabilité des estimations. Le vrai modèle est $Y = X + 0.5 + \varepsilon$, où ε est choisi comme suivant une loi $\mathcal{N}(0, 0.25)$. Nous avons ici 3 répétitions de la mesure de 10 points (x_i, y_i), ou 3 échantillons de taille 10. Le trait en pointillé représente la vraie droite de régression et le trait plein son estimation.

Le statisticien cherche en général à vérifier que les estimateurs utilisés admettent certaines propriétés comme :

– un estimateur $\hat{\beta}$ est-il sans biais ? Par définition $\hat{\beta}$ est sans biais si $\mathbb{E}(\hat{\beta}) = \beta$. En moyenne sur toutes les expériences possibles de taille n, l'estimateur $\hat{\beta}$ moyen sera égal à la valeur inconnue du paramètre. En français, cela signifie qu'en moyenne $\hat{\beta}$ « tombe » sur β ;
– un estimateur $\hat{\beta}$ est-il de variance minimale parmi les estimateurs d'une classe définie ? En d'autres termes, parmi tous les estimateurs de la classe, l'estimateur utilisé admet-il parmi toutes les expériences la plus petite variabilité ?

Pour cela, nous supposons une seconde hypothèse notée $\mathcal{H}_2$ qui s'énonce aussi comme suit : les erreurs sont centrées, de même variance (homoscédasticité) et non corrélées entre elles. Elle permet de calculer les propriétés statistiques des estimateurs. $\mathcal{H}_2 : \mathbb{E}(\varepsilon_i) = 0$, pour $i = 1, \cdots, n$ et $\mathrm{Cov}(\varepsilon_i, \varepsilon_j) = \delta_{ij}\sigma^2$, où $\mathbb{E}(\varepsilon)$ est l'espérance de ε, $\mathrm{Cov}(\varepsilon_i, \varepsilon_j)$ est la covariance entre ε_i et ε_j et $\delta_{ij} = 1$ lorsque $i = j$ et $\delta_{ij} = 0$ lorsque $i \neq j$. Nous avons la première propriété de ces estimateurs (voir exercice 1.2)

Proposition 1.1 (Biais des estimateurs)
$\hat{\beta}_1$ et $\hat{\beta}_2$ estiment sans biais β_1 et β_2, c'est-à-dire que $\mathbb{E}(\hat{\beta}_1) = \beta_1$ et $\mathbb{E}(\hat{\beta}_2) = \beta_2$.

Les estimateurs $\hat{\beta}_1$ et $\hat{\beta}_2$ sont sans biais, nous allons nous intéresser à leur variance. Afin de montrer que ces estimateurs sont de variances minimales dans leur classe, nous allons d'abord calculer leur variance (voir exercices 1.3, 1.4). C'est l'objet de la prochaine proposition.

Proposition 1.2 (Variances de $\hat{\beta}_1$ et $\hat{\beta}_2$)
Les variances et covariance des estimateurs des paramètres valent :

$$\begin{aligned} \mathrm{V}(\hat{\beta}_2) &= \frac{\sigma^2}{\sum (x_i - \bar{x})^2} \\ \mathrm{V}(\hat{\beta}_1) &= \frac{\sigma^2 \sum x_i^2}{n \sum (x_i - \bar{x})^2} \\ \mathrm{Cov}(\hat{\beta}_1, \hat{\beta}_2) &= -\frac{\sigma^2 \bar{x}}{\sum (x_i - \bar{x})^2}. \end{aligned}$$

Cette proposition nous permet d'envisager la précision de l'estimation en utilisant la variance. Plus la variance est faible, plus l'estimateur sera précis. Pour avoir des variances petites, il faut avoir un numérateur petit et (ou) un dénominateur grand. Les estimateurs seront donc de faibles variances lorsque :
– la variance σ^2 est faible. Cela signifie que la variance de Y est faible et donc les mesures sont proches de la droite à estimer ;
– la quantité $\sum (x_i - \bar{x})^2$ est grande, les mesures x_i doivent être dispersées autour de leur moyenne ;
– la quantité $\sum x_i^2$ ne doit pas être trop grande, les points doivent avoir une faible moyenne en valeur absolue. En effet, nous avons

$$\frac{\sum x_i^2}{\sum (x_i - \bar{x})^2} = \frac{\sum x_i^2 - n\bar{x}^2 + n\bar{x}^2}{\sum (x_i - \bar{x})^2} = 1 + \frac{n\bar{x}^2}{\sum (x_i - \bar{x})^2}.$$

L'équation (1.3) indique que la droite des MC passe par le centre de gravité du nuage $(\bar{x}, \bar{y})$. Supposons $\bar{x}$ positif, alors si nous augmentons la pente, l'ordonnée à l'origine va diminuer et vice versa. Nous retrouvons donc le signe négatif pour la covariance entre $\hat{\beta}_1$ et $\hat{\beta}_2$.
Nous terminons cette partie concernant les propriétés par le théorème de Gauss-Markov qui indique que, parmi tous les estimateurs linéaires sans biais, les estimateurs des MC possèdent la plus petite variance (voir exercice 1.5).

Théorème 1.1 (Gauss-Markov)
Parmi les estimateurs sans biais linéaires en Y, les estimateurs $\hat{\beta}_j$ sont de variance minimale.

1.4.2 Résidus et variance résiduelle

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

Définition 1.2 (Résidus)
Les résidus sont définis par

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

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

Nous avons la propriété suivante (voir exercice 1.6).

Proposition 1.3
Dans un modèle de régression linéaire simple, la somme des résidus est nulle.

Intéressons-nous maintenant à l'estimation de σ^2 et construisons un estimateur sans biais $\hat{\sigma}^2$ (voir exercice 1.7) :

Proposition 1.4 (Estimateur de la variance du bruit)
La statistique $\hat{\sigma}^2 = \sum_{i=1}^{n} \hat{\varepsilon}_i^2/(n-2)$ est un estimateur sans biais de σ^2.

1.4.3 Prévision

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

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

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

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

En utilisant la notation $\hat{y}^p_{n+1}$ nous souhaitons insister sur la notion de prévision : la valeur pour laquelle nous effectuons la prévision, ici la $(n+1)^{\text{e}}$, n'a pas servi dans le calcul des estimateurs. Remarquons que cette quantité sera différente de la valeur ajustée, notée $\hat{y}_i$, qui elle fait intervenir la i^{e} observation.
Deux types d'erreurs vont entacher notre prévision, l'une due à la non-connaissance de ε_{n+1} et l'autre due à l'estimation des paramètres.

Proposition 1.5 (Variance de la prévision $\hat{y}^p_{n+1}$)
La variance de la valeur prévue de $\hat{y}^p_{n+1}$ vaut

$$\mathrm{V}\left(\hat{y}^p_{n+1}\right) = \sigma^2\left(\frac{1}{n} + \frac{(x_{n+1}-\bar{x})^2}{\sum(x_i-\bar{x})^2}\right).$$

La variance de $\hat{y}^p_{n+1}$ (voir exercice 1.8) nous donne une idée de la stabilité de l'estimation. En prévision, on s'intéresse généralement à l'erreur que l'on commet entre la vraie valeur à prévoir y_{n+1} et celle que l'on prévoit $\hat{y}^p_{n+1}$. L'erreur peut être simplement résumée par la différence entre ces deux valeurs, c'est ce que nous appellerons l'erreur de prévision. Cette erreur de prévision permet de quantifier la capacité du modèle à prévoir. Nous avons sur ce thème la proposition suivante (voir exercice 1.8).

Proposition 1.6 (Erreur de prévision)
L'erreur de prévision, définie par $\hat{\varepsilon}^p_{n+1} = y_{n+1} - \hat{y}^p_{n+1}$ satisfait les propriétés suivantes :

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

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

1.5 Interprétations géométriques

1.5.1 Représentation des individus

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

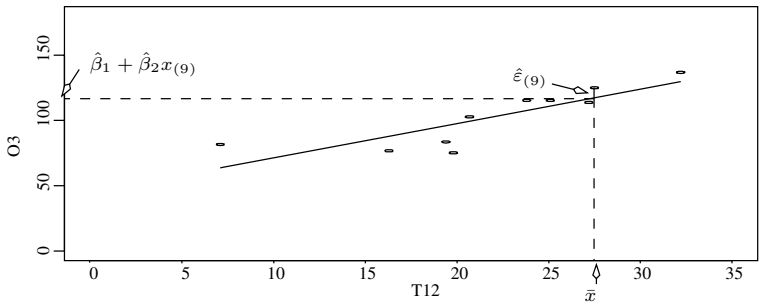

Fig. 1.10 – Représentation des individus.

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

1.5.2 Représentation des variables

Nous pouvons voir le problème d'une autre façon. Nous mesurons n couples de points (x_i, y_i). La variable X et la variable Y peuvent être considérées comme deux vecteurs possédant n coordonnées. Le vecteur X (respectivement Y) admet pour coordonnées les observations $x_1, x_2, \ldots, x_n$ (respectivement $y_1, y_2, \ldots, y_n$). Ces deux vecteurs d'observations appartiennent au même espace $\mathbb{R}^n$: l'espace des variables. Nous pouvons donc représenter les données dans l'espace des variables. Le vecteur $\mathbb{1}$ est également un vecteur de $\mathbb{R}^n$ dont toutes les composantes valent 1. Les 2 vecteurs $\mathbb{1}$ et X engendrent un sous-espace de $\mathbb{R}^n$ de dimension 2. Nous avons supposé que $\mathbb{1}$ et X ne sont pas colinéaires grâce à $\mathcal{H}_1$ mais ces vecteurs ne sont pas obligatoirement orthogonaux. Ces vecteurs sont orthogonaux lorsque $\bar{x}$, la moyenne des observations $x_1, x_2, \ldots, x_n$ vaut zéro.
La régression linéaire peut être vue comme la projection orthogonale du vecteur Y dans le sous-espace de $\mathbb{R}^n$ engendré par $\mathbb{1}$ et X, noté $\Im(X)$ (voir fig. 1.11).

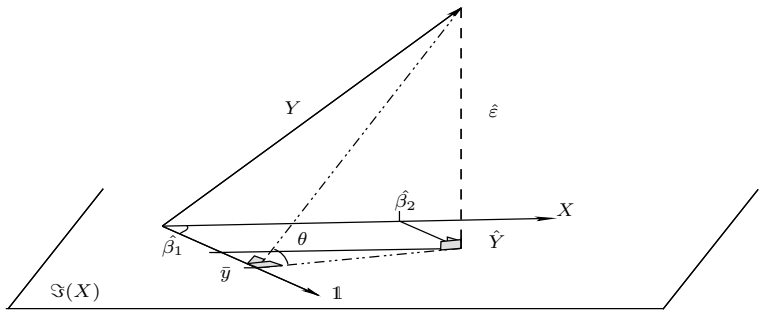

Fig. 1.11 – Représentation de la projection dans l'espace des variables.

Les coefficients $\hat{\beta}_1$ et $\hat{\beta}_2$ s'interprètent comme les composantes de la projection orthogonale notée $\hat{Y}$ de Y sur ce sous-espace.

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

Un modèle, que l'on qualifiera de bon, possédera des estimations $\hat{y}_i$ proches des vraies valeurs y_i. Sur la représentation dans l'espace des variables (fig. 1.11) la qualité peut être évaluée par l'angle θ. Cet angle est compris entre -90 degrés et 90 degrés. Un angle proche de -90 degrés ou de 90 degrés indique un modèle de mauvaise qualité. Le cosinus carré de θ est donc une mesure possible de la qualité du modèle et cette mesure varie entre 0 et 1.
Le théorème de Pythagore nous donne directement que

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

où SCT (respectivement SCE et SCR) représente la somme des carrés totale (respectivement expliquée par le modèle et résiduelle).

Le coefficient de détermination R^2 est défini par

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

c'est-à-dire la part de la variabilité expliquée par le modèle sur la variabilité totale. De nombreux logiciels multiplient cette valeur par 100 afin de donner un pourcentage.

Remarques
Dans ce cas précis, R^2 est le carré du coefficient de corrélation empirique entre les x_i et les y_i et
– le R^2 correspond au cosinus carré de l'angle θ ;
– si $\mathrm{R}^2 = 1$, le modèle explique tout, l'angle θ vaut donc zéro, Y est dans $\Im(X)$ c'est-à-dire que $y_i = \beta_1 + \beta_2 x_i$;

– si $R^2 = 0$, cela veut dire que $\sum(\hat{y}_i - \bar{y})^2 = 0$ et donc que $\hat{y}_i = \bar{y}$. Le modèle de régression linéaire est inadapté ;
– si R^2 est proche de zéro, cela veut dire que Y est quasiment dans l'orthogonal de $\Im(X)$, le modèle de régression linéaire est inadapté, la variable X utilisée n'explique pas la variable Y.

1.6 Inférence statistique

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

$$\mathcal{H}_3 \left\{ \begin{array}{ll} \varepsilon_i & \sim \mathcal{N}(0, \sigma^2) \\ \varepsilon_i & \text{sont indépendants} \end{array} \right.$$

où $\mathcal{N}(0, \sigma^2)$ est une loi normale d'espérance nulle et de variance σ^2. Le modèle de régression devient le modèle paramétrique $\left\{\mathbb{R}^n, \mathcal{B}_{\mathbb{R}^n}, \mathcal{N}(\beta_1 + \beta_2 x, \sigma^2)\right\}$, où β_1, β_2, σ^2 sont à valeurs dans $\mathbb{R}$, $\mathbb{R}$ et $\mathbb{R}^+$ respectivement. La loi des ε_i étant connue, nous en déduisons la loi des y_i. Toutes les preuves de cette section seront détaillées au chapitre 3.
Nous allons envisager dans cette section les propriétés supplémentaires des estimateurs qui découlent de l'hypothèse $\mathcal{H}_3$ (normalité et indépendance des erreurs) :
– lois des estimateurs $\hat{\beta}_1$, $\hat{\beta}_2$ et $\hat{\sigma}^2$;
– intervalles de confiance univariés et bivariés ;
– loi des valeurs prévues $\hat{y}^p_{n+1}$ et intervalle de confiance.
Cette partie est plus technique que les parties précédentes. Afin de faciliter la lecture, considérons les notations suivantes :

$$\begin{array}{rclrcl} \sigma^2_{\hat{\beta}_1} & = & \sigma^2 \left(\dfrac{\sum x_i^2}{n \sum (x_i - \bar{x})^2} \right), & \hat{\sigma}^2_{\hat{\beta}_1} & = & \hat{\sigma}^2 \left(\dfrac{\sum x_i^2}{n \sum (x_i - \bar{x})^2} \right), \\ \sigma^2_{\hat{\beta}_2} & = & \dfrac{\sigma^2}{\sum (x_i - \bar{x})^2}, & \hat{\sigma}^2_{\hat{\beta}_2} & = & \dfrac{\hat{\sigma}^2}{\sum (x_i - \bar{x})^2}, \end{array}$$

où $\hat{\sigma}^2 = \sum \hat{\varepsilon}_i^2/(n-2)$. Cet estimateur est donné au théorème 1.4. Notons que les estimateurs de la colonne de gauche ne sont pas réellement des estimateurs. En effet puisque σ^2 est inconnu, ces estimateurs ne sont pas calculables avec les données. Cependant, ce sont eux qui interviennent dans les lois des estimateurs $\hat{\beta}_1$ et $\hat{\beta}_2$ (voir proposition 1.7). Les estimateurs donnés dans la colonne de droite sont ceux qui sont utilisés (et utilisables) et ils consistent simplement à remplacer σ^2 par $\hat{\sigma}^2$. Les lois des estimateurs sont données dans la proposition suivante.

Proposition 1.7 (Lois des estimateurs : variance connue)
Les lois des estimateurs des MC sont :
(i) $\hat{\beta}_j \sim \mathcal{N}\left(\beta_j, \sigma^2_{\hat{\beta}_j}\right)$ *pour* $j = 1, 2$.
(iii) $\hat{\beta} = \begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix} \sim \mathcal{N}\left(\beta, \sigma^2 V\right)$, $\beta = \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix}$ *et* $V = \dfrac{1}{\sum (x_i - \bar{x})^2} \begin{bmatrix} \sum x_i^2/n & -\bar{x} \\ -\bar{x} & 1 \end{bmatrix}$.
(iv) $\dfrac{(n-2)}{\sigma^2}\hat{\sigma}^2$ *suit une loi du* χ^2 *à* $(n-2)$ *degrés de liberté (ddl)* (χ^2_{n-2}).
(v) $(\hat{\beta}_1, \hat{\beta}_2)$ *et* $\hat{\sigma}^2$ *sont indépendants.*

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

Proposition 1.8 (Lois des estimateurs : variance estimée)
Lorsque σ^2 *est estimée par* $\hat{\sigma}^2$*, nous avons*
(i) pour $j = 1, 2$ $\dfrac{\hat{\beta}_j - \beta_j}{\hat{\sigma}_{\hat{\beta}_j}} \sim \mathcal{T}_{n-2}$ *où* $\mathcal{T}_{n-2}$ *est une loi de Student à* $(n-2)$ *ddl.*
(ii) $\dfrac{1}{2\,\hat{\sigma}^2}(\hat{\beta} - \beta)' V^{-1} (\hat{\beta} - \beta) \sim \mathcal{F}_{2,n-2}$*, où* $\mathcal{F}_{2,n-2}$ *est une loi de Fisher à 2 ddl au numérateur et* $(n-2)$ *ddl au dénominateur.*

Ces dernières propriétés nous permettent de donner des intervalles de confiance (IC) ou des régions de confiance (RC) des paramètres inconnus. En effet, la valeur ponctuelle d'un estimateur est en général insuffisante et il est nécessaire de lui adjoindre un intervalle de confiance. Nous parlerons d'IC quand un paramètre est univarié et de RC quand le paramètre est multivarié.

Proposition 1.9 (IC et RC de niveau $1-\alpha$ pour les paramètres)
(i) Un IC *bilatéral de* β_j *(*$j \in \{1, 2\}$*) est donné par :*

$$\left[\hat{\beta}_j - t_{n-2}(1-\alpha/2)\hat{\sigma}_{\hat{\beta}_j}, \hat{\beta}_j + t_{n-2}(1-\alpha/2)\hat{\sigma}_{\hat{\beta}_j}\right] \tag{1.5}$$

où $t_{n-2}(1-\alpha/2)$ *représente le fractile de niveau* $(1-\alpha/2)$ *d'une loi* $\mathcal{T}_{n-2}$.
(ii) Une RC des deux paramètres inconnus β *est donnée par l'équation suivante :*

$$\frac{1}{2\hat{\sigma}^2}\left[n(\hat{\beta}_1 - \beta_1)^2 + 2n\bar{x}(\hat{\beta}_1 - \beta_1)(\hat{\beta}_2 - \beta_2) + \sum x_i^2 (\hat{\beta}_2 - \beta_2)^2\right] \leq f_{(2,n-2)}(1-\alpha),$$

où $f_{(2,n-2)}(1-\alpha)$ *représente le fractile de niveau* $(1-\alpha)$ *d'une loi de Fisher à* $(2, n-2)$ *ddl.*

(iii) Un IC *de* σ^2 *est donné par :*

$$\left[\frac{(n-2)\hat{\sigma}^2}{c_{n-2}(1-\alpha/2)}, \frac{(n-2)\hat{\sigma}^2}{c_{n-2}(\alpha/2)}\right],$$

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

Remarque
La propriété (ii) donne la RC simultanée des paramètres de la régression $\beta = (\beta_1, \beta_2)'$, appelée ellipse de confiance grâce à la loi du couple. Au contraire (i) donne l'IC d'un paramètre sans tenir compte de la corrélation entre $\hat{\beta}_1$ et $\hat{\beta}_2$. Il est donc délicat de donner une RC du vecteur (β_1, β_2) en juxtaposant les deux IC.

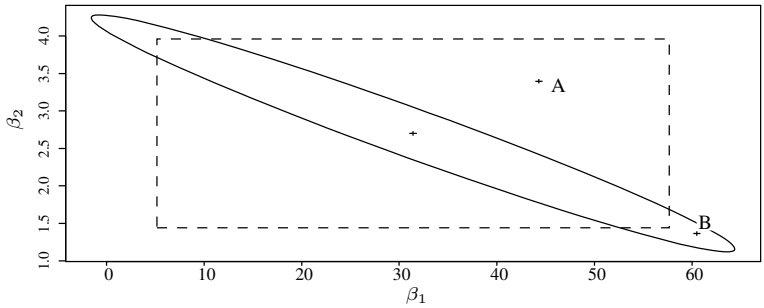

Fig. 1.12 – Comparaison entre ellipse et rectangle de confiance.

Un point peut avoir chaque coordonnée dans son IC respectif mais ne pas appartenir à l'ellipse de confiance. Le point A est un exemple de ce type de point. *A contrario*, un point peut appartenir à la RC sans qu'aucune de ses coordonnées n'appartienne à son IC respectif (le point B). L'ellipse de confiance n'est pas toujours calculée par les logiciels de statistique. Le rectangle de confiance obtenu en juxtaposant les deux intervalles de confiance peut être une bonne approximation de l'ellipse si la corrélation entre $\hat{\beta}_1$ et $\hat{\beta}_2$ est faible.

Nous pouvons donner un intervalle de confiance de la droite de régression.

Proposition 1.10 (IC pour $\mathbb{E}(y_i)$**)**
Un IC *de* $\mathbb{E}(y_i) = \beta_1 + \beta_2 x_i$ *est donné par :*

$$\left[\hat{y}_i \pm t_{n-2}(1-\alpha/2)\hat{\sigma}\sqrt{\frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum(x_l - \bar{x})^2}}\right]. \tag{1.6}$$

En calculant les IC pour tous les points de la droite, nous obtenons une hyperbole de confiance. En effet, lorsque x_i est proche de $\bar{x}$, le terme dominant de la variance est $1/n$, mais dès que x_i s'éloigne de $\bar{x}$, le terme dominant est le terme au carré. Nous avons les mêmes résultats que ceux obtenus à la section (1.4.3). Enonçons le résultat permettant de calculer un intervalle de confiance pour une valeur prévue :

Proposition 1.11 (IC pour y_{n+1})
Un IC *de y_{n+1} est donné par :*

$$\left[\hat{y}^p_{n+1} \pm t_{n-2}(1-\alpha/2)\hat{\sigma}\sqrt{1+\frac{1}{n}+\frac{(x_{n+1}-\bar{x})^2}{\sum(x_i-\bar{x})^2}}\right]. \tag{1.7}$$

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

1.7 Exemples

La concentration en ozone

Nous allons traiter les 50 données journalières de concentration en ozone. La variable à expliquer est la concentration en ozone notée `O3` et la variable explicative est la température notée `T12`.

- Nous commençons par représenter les données.

```
> ozone <- read.table("ozone_simple.txt",header=T,sep=";")
> plot(O3~T12,data=ozone,xlab="T12",ylab="O3")
```

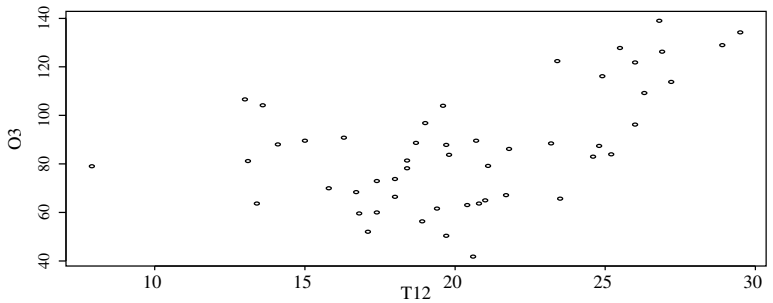

Fig. 1.13 – 50 données journalières de `T12` et `O3`.

Ce graphique permet de vérifier visuellement si une régression linéaire est pertinente. Autrement dit, il suffit de regarder si le nuage de point s'étire le long d'une droite. Bien qu'ici il semble que le nuage s'étire sur une première droite jusqu'à 22 ou 23 degrés C puis selon une autre droite pour les hautes valeurs de températures, nous pouvons tenter une régression linéaire simple.

- Nous effectuons ensuite la régression linéaire, c'est-à-dire la phase d'estimation.

```
> reg <- lm(O3~T12,data=ozone)
```

Afin de consulter les résultats, nous effectuons

```
> summary(reg)
Call:
lm(formula = O3 ~ T12)
Residuals:
    Min      1Q  Median      3Q     Max
-45.256 -15.326  -3.461  17.634  40.072

Coefficients
             Estimate    Std. Error    t value    Pr(>|t|)
(Intercept)    31.4150      13.0584      2.406     0.0200       *
T12             2.7010       0.6266      4.311     8.04e-05   ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 20.5 on 48 degrees of freedom
Multiple R-Squared: 0.2791,     Adjusted R-squared: 0.2641
F-statistic: 18.58 on 1 and 48 DF,  p-value: 8.041e-05
```

Les sorties du logiciel donnent une matrice (sous le mot `Coefficients`) qui comporte pour chaque paramètre (chaque ligne) 5 colonnes. La première colonne contient les estimations des paramètres (colonne `Estimate`), la seconde les écarts-types estimés des paramètres (`Std. Error`). Dans la troisième colonne (`t value`) figure la valeur observée de la statistique de test d'hypothèse $H_0 : \beta_i = 0$ contre $H_1 : \beta_i \neq 0$. La quatrième colonne (`Pr(>|t|)`) contient la probabilité critique (ou « p-value ») qui est la probabilité, pour la statistique de test sous H_0, de dépasser la valeur estimée. Enfin la dernière colonne est une version graphique du test : `***` signifie que le test rejette H_0 pour des erreurs de première espèce supérieures ou égales à 0.001, `**` signifie que le test rejette H_0 pour des erreurs de première espèce supérieures ou égales à 0.01, `*` signifie que le test rejette H_0 pour des erreurs de première espèce supérieures ou égales à 0.05, `.` signifie que le test rejette H_0 pour des erreurs de première espèce supérieures ou égales à 0.1.
Nous rejetons l'hypothèse H_0 pour les deux paramètres estimés au niveau $\alpha = 5\ \%$. Dans le cadre de la régression simple, cela permet d'effectuer de manière rapide un choix de variable pertinente. En toute rigueur, si pour les deux paramètres l'hypothèse H_0 est acceptée, il est nécessaire de reprendre un modèle en supprimant le paramètre dont la probabilité critique est la plus proche de 1. Dans ce cas-là, dès la phase de représentation des données, de gros doutes doivent apparaître sur l'intérêt de la régression linéaire simple.
Le résumé de l'étape d'estimation fait figurer l'estimation de σ qui vaut ici 20.5 ainsi que le nombre $n - 2 = 48$ qui est le nombre de degrés de liberté associés, par exemple, aux tests d'hypothèse $H_0 : \beta_i = 0$ contre $H_1 : \beta_i \neq 0$.
La valeur du R^2 est également donnée, ainsi que le R^2 ajusté noté R^2_a (voir définition 2.4 p. 39). La valeur du R^2 est faible ($R^2 = 0.28$) et nous retrouvons la

remarque effectuée à propos de la figure (fig. 1.13) : une régression linéaire simple n'est peut-être pas adaptée ici.
La dernière ligne, surtout utile en régression multiple, indique le test entre le modèle utilisé et le modèle n'utilisant que la constante comme variable explicative. Nous reviendrons sur ce test au chapitre 3.

• Afin d'examiner la qualité du modèle et des observations, nous traçons la droite ajustée et les observations. Comme il existe une incertitude dans les estimations, nous traçons aussi un intervalle de confiance de la droite (à 95 %).

```
> plot(O3~T12,data=ozone)
> T12 <- seq(min(ozone[,"T12"]),max(ozone[,"T12"]),length=100)
> grille <- data.frame(T12)
> ICdte <- predict(reg,new=grille,interval="confidence",level=0.95)
> matlines(grille$T12,cbind(ICdte),lty=c(1,2,2),col=1)
```

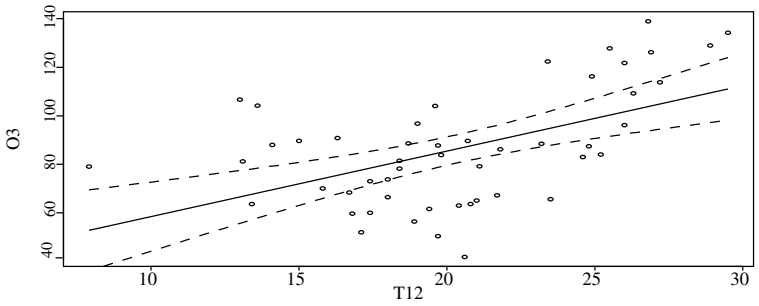

Fig. 1.14 – 50 données journalières de `T12` et `O3` et l'ajustement linéaire obtenu.

Ce graphique permet de vérifier visuellement si une régression est correcte, c'est-à-dire d'analyser la qualité d'ajustement du modèle. Nous constatons que les observations qui possèdent de faibles valeurs ou de fortes valeurs de température sont au-dessus de la droite ajustée (fig. 1.14) alors que les observations qui possèdent des valeurs moyennes sont en dessous. Les erreurs ne semblent donc pas identiquement distribuées. Pour s'en assurer il est aussi possible de tracer les résidus.
Enfin l'intervalle de confiance à 95 % est éloigné de la droite. Cet intervalle peut être vu comme « le modèle peut être n'importe quelle droite dans cette bande ». Il en découle que la qualité de l'estimation ne semble pas être très bonne.

• Dans une optique de prévision, il est nécessaire de s'intéresser à la qualité de prévision. Cette qualité peut être envisagée de manière succincte grâce à l'intervalle de confiance des prévisions. Afin de bien le distinguer de celui de la droite, nous figurons les deux sur le même graphique.

```
> plot(O3~T12,data=ozone,ylim=c(0,150))
> T12 <- seq(min(ozone[,"T12"]),max(ozone[,"T12"]),length=100)
> grille <- data.frame(T12)
```

```
> ICdte <- predict(reg,new=grille,interval="conf",level=0.95)
> ICprev <- predict(reg,new=grille,interval="pred",level=0.95)
> matlines(T12,cbind(ICdte,ICprev[,-1]),lty=c(1,2,2,3,3),col=1)
> legend("topleft",lty=2:3,c("Y","E(Y)"))
```

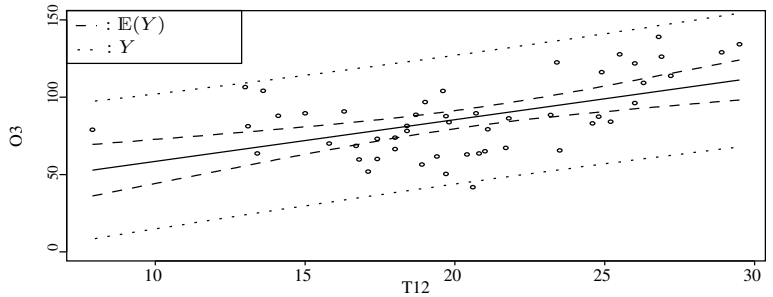

Fig. 1.15 – Droite de régression et intervalles de confiance pour Y et pour $\mathbb{E}(Y)$.

Afin d'illustrer les équations des intervalles de confiance pour les prévisions et la droite ajustée (équations (1.6) et (1.7), p. 20), nous remarquons bien évidemment que l'intervalle de confiance des prévisions est plus grand que l'intervalle de confiance de la droite de régression. L'intervalle de confiance de la droite de régression admet une forme hyperbolique.

• Si nous nous intéressons au rôle des variables, nous pouvons calculer les intervalles de confiance des paramètres *via* la fonction `confint`. Par défaut, le niveau est fixé à 95 %.

```
> IC <- confint(reg,level=0.95)
> IC
                2.5 %    97.5 %
(Intercept) 5.159232 57.67071
T12         1.441180  3.96089
```

L'IC à 95 % sur l'ordonnée à l'origine est étendu (52.5). Cela provient des erreurs (l'estimateur de σ vaut 20.5), mais surtout du fait que les températures sont en moyenne très loin de 0. Cependant, ce coefficient ne fait pas très souvent l'objet d'interprétation.
L'autre IC à 95 % est moins étendu (2.5). Nous constatons qu'il semble exister un effet de la température sur les pics d'ozone, bien que l'on se pose la question de la validité de l'hypothèse linéaire, et donc de la conclusion énoncée ci-dessus.

• Il est conseillé de tracer la région de confiance simultanée des deux paramètres et de comparer cette région aux intervalles de confiance obtenus avec le même degré de confiance. Cette comparaison illustre uniquement la différence entre intervalle simple et région de confiance. En général, l'utilisateur de la méthode choisit l'une ou l'autre forme. Pour cette comparaison, nous utilisons les commandes suivantes :

```
> library(ellipse)
> plot(ellipse(reg,level=0.95),type="l",xlab="",ylab="")
> points(coef(reg)[1], coef(reg)[2],pch=3)
> lines(IC[1,c(1,1,2,2,1)],IC[2,c(1,2,2,1,1)],lty=2)
```

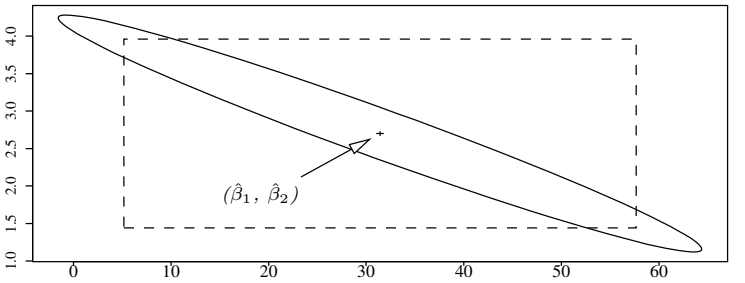

Fig. 1.16 – Région de confiance simultanée des deux paramètres.

Les axes de l'ellipse ne sont pas parallèles aux axes du graphique, les deux estimateurs sont corrélés. Nous retrouvons que la corrélation entre les deux estimateurs est toujours négative (ou nulle), le grand axe de l'ellipse ayant une pente négative. Nous observons bien sûr une différence entre le rectangle de confiance, juxtaposition des deux intervalles de confiance et l'ellipse.

La hauteur des eucalyptus

Nous allons modéliser la hauteur des arbres en fonction de leur circonférence.

- Nous commençons par représenter les données.

```
> eucalypt <- read.table("eucalyptus.txt",header=T,sep=";")
> plot(ht~circ,data=eucalypt,xlab="circ",ylab="ht")
```

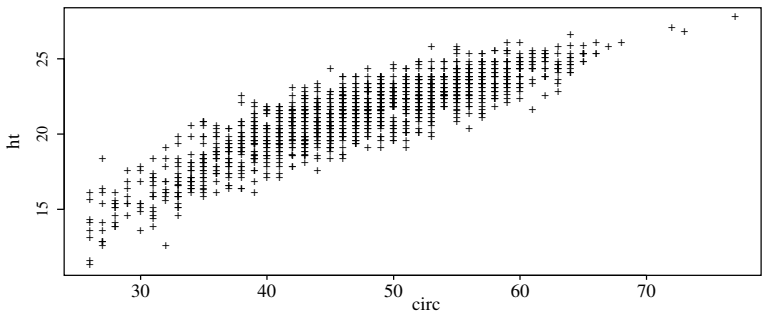

Fig. 1.17 – Représentation des mesures pour les $n = 1429$ eucalyptus mesurés.

Une régression simple semble indiquée, les points étant disposés grossièrement le long d'une droite. Trois arbres ont des circonférences élevées supérieures à 70 cm.

• Nous effectuons ensuite la régression linéaire, c'est-à-dire la phase d'estimation.

```
> reg <- lm(ht~circ,data=eucalypt)
> summary(reg)
Call:
lm(formula = ht ~ circ, data = eucalypt)

Residuals:
     Min       1Q   Median       3Q      Max
-4.76589 -0.78016  0.05567  0.82708  3.69129

Coefficients:
            Estimate  Std. Error  t value  Pr(>|t|)
(Intercept) 9.037476    0.179802    50.26    <2e-16 ***
circ        0.257138    0.003738    68.79    <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.199 on 1427 degrees of freedom
Multiple R-Squared: 0.7683,     Adjusted R-squared: 0.7682
F-statistic:  4732 on 1 and 1427 DF,  p-value: < 2.2e-16
```

Nous retrouvons comme sortie la matrice des informations sur les coefficients, matrice qui comporte 4 colonnes et autant de lignes que de coefficients (voir 1.7, p. 21). Les tests de nullité des deux coefficients indiquent qu'ils semblent tous deux significativement non nuls (quand l'autre coefficient est fixé à la valeur estimée).
Le résumé de l'étape d'estimation fait figurer l'estimation de σ qui vaut ici 1.199 ainsi que le nombre $n - 2 = 1427$ qui est le nombre de degrés de liberté associés, par exemple, aux tests d'hypothèse $H_0 : \beta_i = 0$ contre $H_1 : \beta_i \neq 0$.
La valeur du R^2 est également donnée, ainsi que le R^2_a. La valeur du R^2 est élevée ($R^2 = 0.7683$) et nous retrouvons la remarque déjà faite (fig. 1.17) : une régression linéaire simple semble adaptée.
Le test F entre le modèle utilisé et le modèle n'utilisant que la constante comme variable explicative indique que la circonférence est explicative et que l'on repousse le modèle n'utilisant que la constante comme variable explicative au profit du modèle de régression simple. Ce test n'est pas très utile ici car il équivaut au test de nullité $H_0 : \beta_2 = 0$ contre $H_1 : \beta_2 \neq 0$. De plus, dès la première étape, nous avions remarqué que les points s'étiraient le long d'une droite dont le coefficient directeur était loin d'être nul.

• Afin d'examiner la qualité du modèle et des observations, nous traçons la droite ajustée et les observations. Comme il existe une incertitude dans les estimations, nous traçons aussi un intervalle de confiance de la droite (à 95 %).

```
> plot(ht~circ,data=eucalypt,pch="+",col="grey60")
> grille <- data.frame(circ=seq(min(eucalypt[,"circ"]),
```

```
+                       max(eucalypt[,"circ"]),length=100))
> ICdte <- predict(reg,new=grille,interval="confi",level=0.95)
> matlines(grille$circ,ICdte, lty=c(1,2,2),col=1)
```

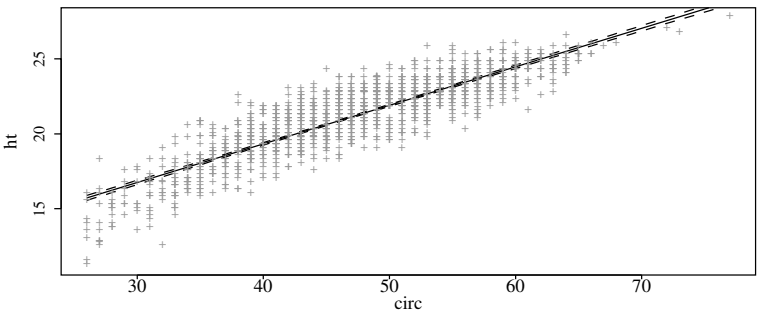

Fig. 1.18 – Données de circonférence/hauteur et ajustement linéaire obtenu.

Ce graphique permet de vérifier visuellement si une régression est correcte, c'est-à-dire de constater la qualité d'ajustement de notre modèle. Nous constatons que les observations sont globalement bien ajustées par le modèle, mais les faibles valeurs de circonférences semblent en majorité situées en dessous de la courbe. Ceci indique qu'un remplacement de cette droite par une courbe serait une amélioration possible. Peut-être qu'un modèle de régression simple du type

$$\texttt{ht} \quad = \quad \beta_0 + \beta_1 \sqrt{\texttt{circ}} + \varepsilon,$$

serait plus adapté. Remarquons aussi que les 3 circonférences les plus fortes (supérieures à 70 cm) sont bien ajustées par le modèle. Ces 3 individus sont donc différents en terme de circonférence mais bien ajustés par le modèle.
Enfin, l'intervalle de confiance à 95 % est proche de la droite. Cet intervalle peut être vu comme « le modèle peut être n'importe quelle droite dans cette bande ». Il en découle que la qualité de l'estimation semble être très bonne, ce qui est normal car le nombre d'individus (i.e. le nombre d'arbres) est très élevé et les données sont bien réparties le long d'une droite.

• Dans une optique de prévision, il est nécessaire de s'intéresser à la qualité de prévision. Cette qualité peut être envisagée de manière succincte grâce aux intervalles de confiance, de la droite ajustée et des prévisions.

```
> plot(ht~circ,data=eucalypt,pch="+",col="grey60")
> circ <- seq(min(eucalypt[,"circ"]),max(eucalypt[,"circ"]),len=100)
> grille <- data.frame(circ)
> ICdte <- predict(reg,new=grille,interval="conf",level=0.95)
> ICprev <- predict(reg,new=grille,interval="pred",level=0.95)
> matlines(circ,cbind(ICdte,ICprev[,-1]),lty=c(1,2,2,3,3),col=1)
```

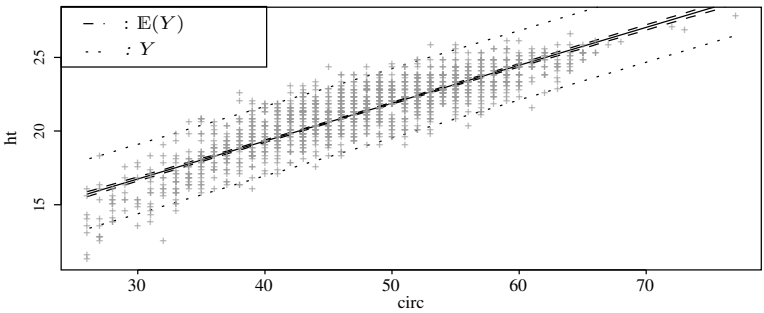

Fig. 1.19 – Droite de régression et intervalles de confiance pour Y et pour $\mathbb{E}(Y)$.

Rien de notable sur l'intervalle de prévision, mis à part le fait qu'il est nécessaire de bien distinguer l'intervalle de confiance de la droite et de la prévision.

1.8 Exercices

Exercice 1.1 (Questions de cours)

1. Lors d'une régression simple, si le R^2 vaut 1, les points sont alignés :
   A. non,
   B. oui,
   C. pas obligatoirement.
2. La droite des MC d'une régression simple passe par le point $(\bar{x}, \bar{y})$:
   A. toujours,
   B. jamais,
   C. parfois.
3. Nous avons effectué une régression simple, nous recevons une nouvelle observation x_N et nous calculons la prévision correspondante $\hat{y}_N$. La variance de la valeur prévue est minimale lorsque
   A. $x_N = 0$,
   B. $x_N = \bar{x}$,
   C. aucun rapport.
4. Le vecteur $\hat{Y}$ est orthogonal au vecteur des résidus estimés $\hat{\varepsilon}$:
   A. toujours,
   B. jamais,
   C. Parfois.

Exercice 1.2 (Biais des estimateurs)
Calculer le biais de $\hat{\beta}_2$ et $\hat{\beta}_1$.

Exercice 1.3 (Variance des estimateurs)
Calculer la variance de $\hat{\beta}_2$ puis la variance de $\hat{\beta}_1$ (indice : calculer $\mathrm{Cov}(\bar{y}, \hat{\beta}_2)$).

Exercice 1.4 (Covariance de $\hat{\beta}_1$ et $\hat{\beta}_2$)
Calculer la covariance entre $\hat{\beta}_1$ et $\hat{\beta}_2$.

Exercice 1.5 (†Théorème de Gauss-Markov)
Démontrer le théorème de Gauss-Markov en posant $\tilde{\beta}_2 = \sum_{i=1}^n \lambda_i y_i$, un estimateur linéaire quelconque (indice : trouver deux conditions sur la somme des λ_i pour que $\tilde{\beta}_2$ ne soit pas biaisé, puis calculer la variance en introduisant $\hat{\beta}_2$).

Exercice 1.6 (Somme des résidus)
Montrer que, dans un modèle de régression linéaire simple, la somme des résidus est nulle.

Exercice 1.7 (Estimateur de la variance du bruit)
Montrer que, dans un modèle de régression linéaire simple, la statistique $\hat{\sigma}^2 = \sum_{i=1}^n \hat{\varepsilon}_i^2/(n-2)$ est un estimateur sans biais de σ^2.

Exercice 1.8 (Prévision)
Calculer la variance de $\hat{y}_{n+1}^p$ puis celle de l'erreur de prévision $\hat{\varepsilon}_{n+1}^p$.

Exercice 1.9 (R^2 et coefficient de corrélation)
Démontrer que le R^2 est égal au carré du coefficient de corrélation empirique entre les x_i et les y_i.

Exercice 1.10 (Les arbres)
Nous souhaitons exprimer la hauteur y d'un arbre d'une essence donnée en fonction de son diamètre x à 1 m 30 du sol. Pour ce faire, nous avons mesuré 20 couples « diamètre-hauteur ». Nous avons effectué les calculs suivants :

$$\bar{x} = 34.9 \qquad \frac{1}{20}\sum_{i=1}^{20}(x_i - \bar{x})^2 = 28.29 \qquad \bar{y} = 18.34$$

$$\frac{1}{20}\sum_{i=1}^{20}(y_i - \bar{y})^2 = 2.85 \qquad \frac{1}{20}\sum_{i=1}^{20}(x_i - \bar{x})(y_i - \bar{y}) = 6.26.$$

1. On note $\hat{y} = \hat{\beta}_1 + \hat{\beta}_2 x$, la droite de régression. Donner l'expression de $\hat{\beta}_2$ en fonction des statistiques élémentaires ci-dessus. Calculer $\hat{\beta}_1$ et $\hat{\beta}_2$.
2. Donner et commenter une mesure de la qualité de l'ajustement des données au modèle. Exprimer cette mesure en fonction des statistiques élémentaires.
3. Cette question traite des tests qui seront vus au chapitre 3. Cependant, cette question peut être résolue grâce à la section exemple. Les estimations des écarts-types de $\hat{\beta}_1$ et de $\hat{\beta}_2$ donnent $\hat{\sigma}_{\hat{\beta}_1} = 1.89$ et $\hat{\sigma}_{\hat{\beta}_2} = 0.05$. Testez $H_0 : \beta_j = 0$ contre $H_1 : \beta_j \neq 0$ pour $j = 0, 1$. Pourquoi ce test est-il intéressant dans notre contexte ? Que pensez-vous du résultat ?

Exercice 1.11 (Modèle quadratique)
Au vu du graphique 1.13, nous souhaitons modéliser l'ozone par la température au carré.

1. Ecrire le modèle et estimer les paramètres.
2. Comparer ce modèle au modèle linéaire classique.