0.1 Introduction

L'été au Québec est particulièrement attendu après un long hiver : on peut profiter de la température élevée pour se faire bronzer et prendre du soleil. Les journées chaudes sont typiquement appréciées pour la baignade.

Pour la logistique des piscines municipales, connaître le nombre de personnes qui pourraient se présenter à la piscine pour une journée donnée est une donnée cruciale : on peut recruter le nombre approprié de sauveteurs, planifier les quantités appropriées de chlore et la gestion approprié de la collecte des déchets sur les terrains publics entourant les piscines.

Ainsi, pour des fins de planification, il est fort utile d'estimer le nombre de personnes qui se présentent quotidiennement à la piscine. *Intuitivement*, on sait que la température influence le nombre de personnes qui veulent se baigner. Il est donc naturel de postuler l'existence d'une relation quantitative entre le nombre de personnes à la piscine (qu'on notera y) et la température d'une journée donnée (qu'on notera x_1). Évidemment, on ne connaît pas encore cette relation. Il est même possible que ce postulat soit faux ! C'est justement le rôle de **l'analyse de régressions** que d'estimer cette analyse et de vérifier si ce postulat est valide.

Pour estimer l'effet de la température sur le nombre de personnes, l'équipe de planification opérationnelle a pris des mesures simultanées de la température et du nombre de personnes à la piscine. Pour une journée i donnée, ils ont mesuré x_{i1}, soit la température en degrés Celsius et y_i, le nombre de personnes à la piscine. L'équipe a donc construit un **échantillon statistique** qui est présenté au tableau 1. On peut remarquer qu'au premier jour, la température était de 24 degrés Celsius et ils ont observé 20 personnes à la piscine. Le deuxième jour, ils ont observé une température de 26 degrés Celsius et 50 personnes se sont présentées à la piscine. Le troisième et dernier jour, ils ont observé une participation de 40 personnes pour une température de 30 degrés Celsius.

Pour simplifier l'exposé de ce texte, l'équipe (fictive) de planification n'a recueilli que trois observations ($n = 3$), ce qui est très peu. Cependant, c'est amplement suffisant pour illustrer les fondements mathématiques de la régression. Plus tard, quand nous aborderons les fondements *statistiques* des régressions, nous arriverons à la conclusion qu'un échantillon de trois observations est de *faible puissance*. Si on devait faire une estimation pratique, il faudrait un échantillon beaucoup plus grand. Nous reviendrons sur les propriétés désirables d'un échantillon (notamment sa taille) ultérieurement. Retenons à ce stade que si l'échantillon était plus grand, la mécanique de calcul exposée dans les pages qui suivent serait essentiellement la même. En fait, le lecteur remarquera que les formules développées ci-dessous tiendrons pour n'importe quelle taille d'échantillon (n'importe quelle valeur de n).

Ceci étant dit, on cherche à estimer une relation linéaire « mystère » liant la température au nombre de personnes à la piscine :

$$y_i = \beta_0 + \beta_1 x_{i1} + e_i. \tag{1}$$

Dans cette expression, trois nouveaux termes méritent des explications :

TABLE 1 – Échantillon recueilli par l'équipe de planification

Observation i	Température x_{i1}	Personnes y_i
1	24	20
2	26	50
3	30	40

1. Les termes β_0 et β_1 sont les paramètres inconnus de la relation linéaire qu'on cherche à estimer. β_0 est l'ordonnée à l'origine et β_1 est la pente. C'est l'équivalent du b et du a dans la relation $y = ax + b$. En statistique, l'usage veut qu'on emploie cependant la lettre « beta » suivi d'un indice (β_0 et β_1).[1] Ces coefficients inconnus seront déduits à partir de l'échantillon pour représenter les données « le mieux possible ».
2. Le terme e_i est une nouvelle variable qui représente **l'erreur d'estimation**. Comme la relation linéaire mystère ne parviendra pas à reproduire parfaitement les données observées, la valeur prédite par l'estimation (la partie $\beta_0 + \beta_1 x_{i1}$) doit être complémentée par une **valeur résiduelle**, ou par une **erreur de prédiction**, pour arriver à la valeur observée. Comme on ne connaît pas encore l'équation permettant de faire des prédictions, on ignore encore quelle sera l'erreur de prédiction de chaque observation (e_1, e_2, e_3). Ces erreurs d'estimation seront déduites lorsqu'on aura identifié les coefficients β_0, β_1. De toute évidence, une propriété désirable du modèle qu'on cherche à établir est que les erreurs de prédiction soient *minimales*.

Les fondements mathématiques d'une régression consistent à trouver les nombres β_0^*, β_1^* qui représentent le mieux les données observées. D'entrée de jeu, je peux donner la réponse associée aux données décrites au tableau 1. L'équation mystère recherchée est donnée par :

$$y_i = -30 + \frac{5}{2} x_{i1} + e_i. \tag{2}$$

Cette prédiction suggère qu'une augmentation de la température de 1 degré Celsius induit une augmentation de 2.5 personnes à la piscine car le coefficient estimé de pente (β_1^*) est égal à 5/2 (2.5). Les prédictions associées à ce modèle et les erreurs de prédiction sont aussi présentées au tableau 4.

Ce texte vise donc à répondre à cette question : comment suis-je arrivé à ces nombres ? Ou peut-être devrais-je être plus honnête : comment le logiciel que j'ai employé pour faire les calculs est arrivé à ces nombres ? Pour répondre à cette question, il faut maîtriser deux ingrédients de base, soit les paraboles multivariées (section 0.2) et la notion de covariance (section 0.3). Les paraboles

1. Bêta est la lettre « b » en grec.

TABLE 2 – Échantillon, valeur prédite et erreur d'estimation

Observation i	Température y_i	Personnes x_{i1}	Valeur prédite $\beta_0^* + \beta_1^* x_{i1}$	Erreur d'estimation e_i^*
1	24	20	30	-10
2	26	50	35	15
3	30	40	45	-5

multivariées ne sont rien d'autre que la généralisation des équations quadratiques univariées ($y = ax^2 + bx + c$) à plus d'une variable ($x_1, x_2, \dots$). La covariance est quant-à-elle un concept statistique qui mesure à quel point deux variables varient (« variance ») conjointement (« co »), c'est-à-dire à quel point elles bougent ensemble.

Quand ces deux notions seront acquises, nous pourrons ensuite voir comment elles sont employées pour calculer les nombres β_0, β_1 et par extension, les erreurs de prédiction (sections 0.4 et 0.5). Il est utile, avant d'entamer l'étude de ces concepts, de revoir les mathématiques associées aux équations quadratiques et aux opérateurs de sommation dans le matériel de révision.

0.2 Paraboles multivariées

Cette section généralise les équations quadratiques à une variable ($y = ax^2 + bx + c$) à deux ou trois variables. Plutôt que d'employer les variables $x_1, x_2, \dots$, j'emploierai la notation d'usage associée aux coefficients de régression (β_0, β_1, etc) car c'est à ces variables que l'ensemble des discussions associées aux paraboles s'appliqueront. Une **forme quadratique** $f(\beta_0, \beta_1)$ à deux variables a la forme suivante :

$$f(\beta_0, \beta_1) = c_1\beta_0^2 + c_2\beta_0 + c_3\beta_0\beta_1 + c_4\beta_1 + c_5\beta_1^2 + c_6, \tag{3}$$

où les c_i sont des nombres connus, équivalents à a, b et c dans la version univariée. Cette forme quadratique a quelques caractéristiques :

1. Comme énoncé, elle dépend de deux variables, soit β_0 et β_1.
2. La plus haute puissance (l'exposant le plus élevé) de chaque variable est deux (2). Qui plus est, la plus haute puissance de chaque produit de variables est 2 (β_0^2, β_1^2 ou encore $\beta_0\beta_1$).
3. Si les coefficients c_3, c_4 et c_5 sont égaux à zéro, ou si les coefficients c_1, c_2, c_3 sont égaux à zéro, on retrouve l'équation quadratique à une variable qui est bien familière.

Graphiquement, on peut représenter cette forme quadratique dans un espace à trois dimensions (voir la Figure 1). C'est la généralisation tri-dimensionnelle de l'équation quadratique univariée (représentée en deux dimensions).

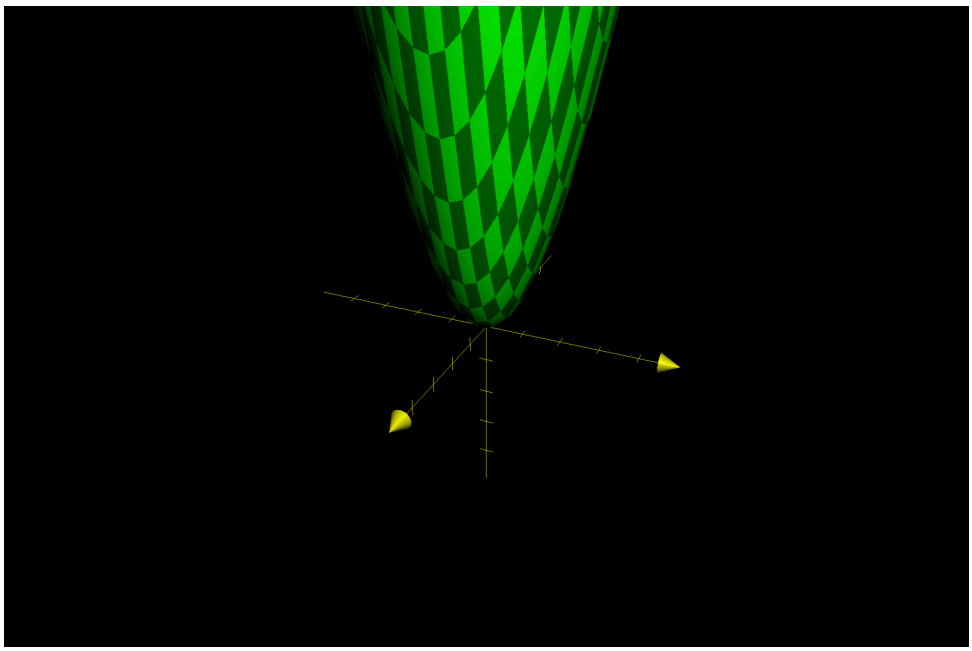

FIGURE 1 – Représentation de $\beta_0^2 + \beta_1^2$ dans un espace tri-dimensionnel

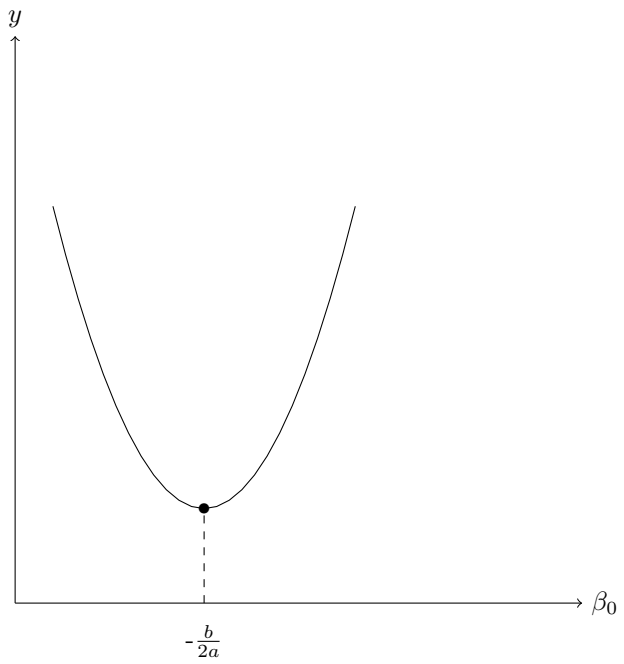

FIGURE 2 – Une parabole univariée $y = a\beta_0^2 + b\beta_0 + c$ prends sa valeur minimale à $\beta_0^* = -\frac{b}{2a}$.

Une équation quadratique à trois variables a quant-à-elle la forme suivante :

$$f(\beta_0, \beta_1, \beta_2) = c_1\beta_0^2 + c_2\beta_0 + c_3\beta_1 + c_4\beta_2 + c_5\beta_0\beta_1 + c_6\beta_0\beta_2 + c_7\beta_1\beta_2 + c_8\beta_1^2 + c_9\beta_2^2 + c_{10}. \quad (4)$$

Outre le nombre de variables, la propriété centrale de la forme quadratique demeure la même que pour le cas à deux variables, à savoir que la plus haute puissance de chaque produit de variable est deux (2) (β_0^2, β_1^2 , β_2^2 ou encore $\beta_0\beta_1$, etc). Contrairement à la forme quadratique à deux variables, on ne peut pas représenter cette forme quadratique graphiquement, car il faudrait visualiser un espace quadridimensionnel.

0.2.1 Trouver le sommet d'une forme quadratique

Notre intérêt pour les formes quadratiques vient du fait qu'on voudra, dans les pages à venir, trouver la valeur minimale que ce type de fonction peut prendre. On cherche le « creux » de la parabole multidimensionnelle. Dans une forme quadratique univarée, on se souvient que cette valeur minimale est donnée à la valeur $\beta_0^* = -\frac{b}{2a}$ (voir la Figure 2) .

Pour une forme quadratique à plusieurs variables, l'intuition est la même. Le minimum de la parabole sera aussi au point spécifié par les coordonnées « $-\frac{b}{2a}$

», mais il faut cependant tenir compte d'une difficulté supplémentaire. En effet, il faut spécifier cette coordonnée pour chacune des dimensions de la parabole $(\beta_0, \beta_1, \dots)$.

Pour trouver les coordonnées selon chaque dimension, l'idée consiste à organiser la forme quadratique en fonction d'une seule variable, comme si c'était une équation univariée. Il faut ensuite identifier les termes équivalents à b et a pour appliquer la formule $-\frac{b}{2a}$ selon cette dimension. Ensuite, on recommence en choisissant une autre variable, ce qui nous donnera une autre équation pour une autre coordonnée. En procédant ainsi pour chaque variable, on obtient alors un système d'équation qui nous permet d'identifier le sommet. Il reste alors à résoudre ce système d'équation avec les méthodes usuelles d'algèbre.

Pour fixer les idées, il est utile de donner un exemple concret : la forme quadratique de l'équation 3 peut s'écrire comme ci-dessous, où sa structure est ré-organisée pour qu'elle prenne le forme d'une parabole univariée en fonction de x_1 :

$$f(\beta_0, \beta_1) = \underbrace{c_1}_{a} \beta_0^2 + \underbrace{(c_2 + c_3\beta_1)}_{b} \beta_0 + \underbrace{\left(c_4\beta_1 + c_5\beta_1^2 + c_6\right)}_{c}. \tag{5}$$

Dans cette mise en évidence, le coefficient analogue à a est donné par c_1, le coefficient analogue à b est donné par $c_2 + c_3\beta_1$ et le coefficient analogue à c est donné par $c_4x_2 + c_5\beta_1^2 + c_6$. Bien que certains de ces coefficients dépendent de β_1, aucun ne dépend de β_0. Du point de vue de la dimension β_0, ces coefficients sont fixes. En conséquence, le sommet en β_0 sera donné par l'équivalent de la formule $-\frac{b}{2a}$, soit :

$$\beta_0 = -\frac{c_2 + c_3\beta_1}{2c_1} \tag{6}$$

En faisant une démarche similaire pour la variable β_1, on trouve que le sommet sera donné à la coordonnée :

$$\beta_1^* = -\frac{c_4 + c_3\beta_0^*}{2c_5} \tag{7}$$

Remarquez que c'est un système à deux équations et deux inconnus en β_0, β_1 qu'on peut résoudre avec les techniques usuelles (voir les mathématiques de révision). Un peu d'algèbre implique la solution suivante :

$$\beta_0^* = -\frac{2c_2c_5 + c_3c_4}{4c_1c_5 - c_3^2}, \tag{8}$$

$$\beta_1^* = -\frac{c_4}{2c_5} + \frac{c_3}{2c_5}\left(\frac{2c_2c_5 + c_3c_4}{4c_1c_5 - c_3^2}\right). \tag{9}$$

Ces deux points nous donnent l'extrémum de la fonction. Dans le cadre de l'analyse de régression, cet extremum est un minimum, soit la plus petite valeur que peut prendre la fonction.

En ce qui concerne une forme quadratique à trois variables, on peut reprendre la même démarche et la réorganiser en fonction de chacune des variables :

$$f(\beta_0, \beta_1, \beta_2\beta_2) = \underbrace{c_1}_{a}\beta_0^2 + \underbrace{(c_2 + c_5\beta_1 + c_6\beta_2)}_{b}\beta_0 + \underbrace{\left(c_3\beta_1 + c_4\beta_2 + c_7\beta_1\beta_2 + c_8\beta_1^2 + c_9\beta_2^2 + c_{10}\right)}_{c}. \quad \text{(Selon } \beta_0\text{)}$$

$$f(\beta_0, \beta_1, \beta_2) = \underbrace{c_8}_{a}\beta_1^2 + \underbrace{(c_3 + c_5\beta_0 + c_7\beta_2)}_{b}\beta_1 + \underbrace{\left(c_1\beta_0^2 + c_2\beta_0 + c_4\beta_2 + c_6\beta_0\beta_2 + c_9\beta_2^2 + c_{10}\right)}_{c}. \quad \text{(Selon } \beta_1\text{)}$$

$$f(\beta_0, \beta_1, \beta_2) = \underbrace{c_9}_{a}\beta_2^2 + \underbrace{(c_4 + c_6\beta_0 + c_7\beta_1)}_{b}\beta_2 + \underbrace{\left(c_1\beta_0^2 + c_2\beta_0 + c_3\beta_1 + c_5x_1\beta_1 + c_8\beta_1^2 + c_{10}\right)}_{c}. \quad \text{(Selon } x_3\text{)}$$

En conséquence, les coordonnées pour le sommet sont données par l'application de la formule $-\frac{b}{2a}$ à chacune de ces équations :

$$\beta_0 = -\frac{c_2 + c_5\beta_1 + c_6\beta_2}{2c_1}, \tag{10}$$

$$\beta_1 = -\frac{c_3 + c_5\beta_0 + c_7\beta_2}{2c_8}, \tag{11}$$

$$\beta_2 = -\frac{c_4 + c_6\beta_0 + c_7\beta_3}{2c_9}, \tag{12}$$

ce qui constitue un système de trois équations à trois inconnues. On peut le résoudre avec les techniques usuelles d'algèbre.

On peut généraliser la démarche identifiée ci-dessus pour trouver le sommet d'une forme quadratique à $k+1$ variables en appliquant une démarche similaire :

1. Regrouper les termes en fonction d'une seule variable et identifier la formule $-\frac{b}{2a}$ associée à cette variable. Répéter pour chaque variable.
2. Cela génèrera un système de $k+1$ équations à $k+1$ inconnues qu'il faut ensuite résoudre avec les techniques usuelles d'algèbre.
3. Ce système d'équations aura pour solution les coordonnées de la valeur minimale que peut prendre la forme quadratique.

0.2.2 Pourquoi faut-il connaître ces formes quadratiques ?

Lorsque viendra le temps de trouver les coefficients de l'équation 1, le critère employé pour trouver ces coefficients sera de les choisir de manière à *minimiser* les erreurs de prédiction. Nous verrons dans quelques pages que les erreurs de prédiction s'expriment comme une forme quadratique en β_0 et β_1. Trouver le minimum sera donc équivalent à trouver le sommet de cette forme quadratique. Connaître les formes quadratiques permet donc de trouver la solution à l'analyse de régression posé en introduction.

TABLE 3 – Échantillon quelconque

Indice d'observation	Variable y	Variable x_1
1	y_1	x_{11}
2	y_2	x_{12}
$\vdots$	$\vdots$	$\vdots$
i	y_i	x_{i1}
$\vdots$	$\vdots$	$\vdots$
$n-1$	y_{n-1}	$x_{(n-1)1}$
n	y_n	x_{n1}

0.2.3 Exercices

Identifiez le minimum de chacune des formes quadratiques à deux variables ci-dessous :

$$f_1(\beta_0, \beta_1\beta_1) = \beta_0^2 + \beta_1^2 \qquad f_2(\beta_0, \beta_1) = \beta_0^2 + 2\beta_0 + \beta_1^2, \tag{13}$$

$$f_3(\beta_0, \beta_1) = 3\beta_0^2 - \beta_1 + \beta_1^2 \qquad f_4(\beta_0, \beta_1) = \beta_0^2 + 2\beta_0\beta_1 + \beta_1 + 2\beta_1^2. \tag{14}$$

0.2.4 Solutions

Les solutions sont données par :

$$\beta_0^* = \beta_1^* = 0, \tag{15}$$

$$\beta_0^* = -1, \qquad \beta_1^* = 0, \tag{16}$$

$$\beta_0^* = 0, \qquad \beta_1^* = \frac{1}{2}, \tag{17}$$

$$\beta_0^* = 0, \qquad \beta_1^* = -\frac{1}{4}. \tag{18}$$

0.3 La covariance

Cette section discute de deux variables statistiques x_1 et y quelconques qui ont été échantillonnées en même temps. Chaque observation de ces variables seront identifiées par un indice i, si bien qu'on parle d'une paire d'observations (x_{i1}, y_i). Cet échantillon quelconque est décrit au tableau 3. Pour rester proche de l'exemple en introduction, on peut penser à la variable x_1 comme la température d'une journée et la variable y comme le nombre de personnes observées à la piscine. Un couple (x_{i1}, y_i) représente donc une observation conjointe de la température et du nombre de personnes pour une journée (observation) donnée.

0.3.1 Rappel de la moyenne et quelques propriétés associées

Pour une variable statistique x_1 quelconque, sa moyenne empirique est donnée par :

$$\bar{x}_1 \overset{\text{def}}{=} \frac{1}{n}\sum_{i=1}^{n} x_{i1}. \tag{19}$$

C'est un nombre qui caractérise la tendance centrale de la variable X dans l'échantillon. Pour employer une analogie physique, si on plaçait un poids sur une planche à la position équivalente à la valeur de chaque observation x_i, la moyenne correspondrait alors à la position sur la planche ou elle est en équilibre. La moyenne mesure le « point d'équilibre » des données.

Deux conséquences liées à la définition de la moyenne seront utiles à la section 0.4, si bien que je les introduit immédiatement. La première est une simple conséquence de la multiplication de la définition par n de chaque côté de l'égalité :

$$n\bar{x}_1 = \sum_{i=1}^{n} x_{i1},. \tag{20}$$

Cette reformulation de la définition est utile pour simplifier des expressions, car le terme $\sum_i x_{1i}$ apparaît fréquemment, si bien qu'on peut le remplacer par $n\bar{x}_1$.

La deuxième conséquence doit cependant se justifier avec un peu de développements algébriques. Je montre la formule tout de suite et je fais ensuite son developpement :

$$\sum_i (x_{i1} - \bar{x}_1) = 0. \tag{21}$$

En français, cette expression stipule que les déviations des observations par rapport à la moyenne $(x_{i1} - \bar{x}_1)$ s'annulent lorsqu'on les additionne ensemble. Pour le voir, on peut faire le développement suivant :

$$\begin{aligned}
\sum_{i=1}^{n}(x_{i1} - \bar{x}_1) &= \sum_{i=1}^{n} x_{i1} - \sum_{i=1}^{n}\bar{x}_1, && \text{(distribution de la sommation)}\\
&= n\bar{x}_1 - \sum_{i=1}^{n}\bar{x}_1, && \text{(application de l'équation 20)}\\
&= n\bar{x}_1 - n\bar{x}_1, && \text{(sommation sur une constante)}\\
&= 0.
\end{aligned}$$

En particulier, la multiplication de cette formule par une constante ne changera pas l'égalité. Par exemple, en la multipliant par $\bar{y}$, nous aurons :

$$\bar{y}\sum_i (x_{i1} - \bar{x}_1) = \sum_i \bar{y}(x_{i1} - \bar{x}_1) = 0. \tag{22}$$

Pour l'instant, ces deux propriétés de la moyenne (équations 20 et 22) n'ont pas encore de sens, mais nous les emploierons plus tard (soit à l'équation 41 ci-dessous).

0.3.2 La covariance

L'étymologie du mot covariance révèle beaucoup sur son sens. Le terme « co » signifie « ensemble » ou « conjointement » alors que le terme variance signifie « bouger » ou « changement ». En somme, la covariance mesure à quel point deux variables statistiques « bougent ensemble ».

La covariance est une mesure quantitative dont le signe révèle la direction du changement entre les variables. Un signe positif signifie que lorsqu'une variable augmente, l'autre variable statistique a aussi tendance à augmenter. Inversement, une covariance négative signifie que lorsque'une variable augmente, l'autre variable a tendance à diminuer. Finalement, une covariance de magnitude élevée indique qu'un changement modeste d'une variable implique implique un grand changement de l'autre variable.

Dans l'exemple mentionné en introduction, on sait que lorsque la température (t) augmente, le nombre de personnes à la piscine p a aussi tendance à augmenter. Intuitivement, on peut donc s'attendre à ce que la covariance entre ces deux variables soit positive. Si on mesurait au contraire la relation entre les ventes de billet de cinéma et la température, on pourrait s'attendre à une covariance négative. En effet, si la température est élevée, on peut présumer que les gens préfèreront être à l'extérieur (peut-être à la piscine !) plutôt que dans une salle de cinéma.

Pour deux variables statistiques x_1 et y, la covariance est donnée par la formule :

$$\text{cov}(x_1, y) \overset{\text{def}}{=} \frac{1}{n-1} \sum_{i=1}^{n} (x_{i1} - \bar{x}_1)(y_i - \bar{y}), \tag{23}$$

où $\bar{x}_1$ et $\bar{y}$ sont respectivement la moyenne empirique de la variable x_1 et de la variable y.

Outre la description intuitive fournie dans les deux premiers paragraphes, comment interpréter cette formule ? La meilleure manière de développer l'intuition consiste à analyser l'échantillon dans un plan cartésien centré autour de $\bar{x}_1$ et de $\bar{y}$. Je présente ce plan cartésien à la Figure 3. Les barres pointillées représentent les droites $y = \bar{y}$ et $x = \bar{x}_1$. L'intersection des deux barres pointillées est donc le point équivalent à la moyenne des deux variables $(\bar{x}_1, \bar{y})$. Ces lignes pointillées séparent le plan cartésien en quatre sous-espaces, numérotés de un (1) à quatre (4). Chaque sous-espace est aussi identifié d'un signe (+ ou -), que j'explique ci-dessous.

La compréhension de l'analyse de la formule de covariance commence par l'analyse d'un point (x_{i1}, y_i) quelconque qui se trouve dans le sous-espace # 1. On peut remarquer que le point x_i se trouve à gauche de la moyenne $\bar{x}_1$, ce qui veut dire que la moyenne est plus élevée que l'observation. En conséquence,

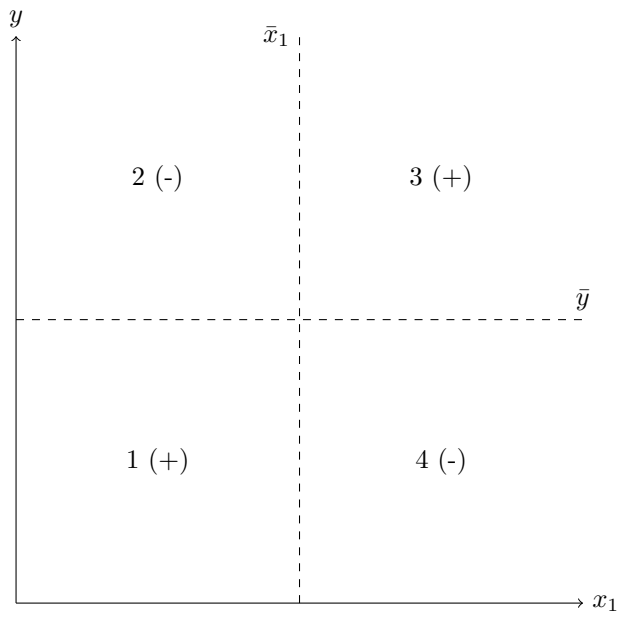

FIGURE 3 – Espace graphique d'analyse de la covariance.

le terme $(x_{i1} - \bar{x}_1)$ est négatif. Similairement, le point y_i est inférieur au point $\bar{y}$ si bien que le terme $(y_i - \bar{y})$ est aussi négatif. Il s'en suit que le produit des deux termes $(x_{i1} - \bar{x}_1)(y_i - \bar{y})$ est positif car deux négatifs multipliés ensembles donnent un nombre positif. La contribution de ce point à la sommation générale est donc positive. Comme on a pris un point quelconque dans le sous-espace, on devrait conclure que *n'importe quel* point qui se retrouve dans ce sous-espace #1 fournira aussi une contribution positive à la covariance. En termes de signe, tous les points dans ce sous-espace ajouteront une valeur positive à la covariance. C'est pourquoi le sous-espace est identifié avec un signe « + ».

Les points se retrouvant dans le deuxième sous-espace fourniront quant-à-eux une contribution négative à la covariance. En effet, le point y_i sera supérieur à $\bar{y}$, si bien que le terme $(y_i - \bar{y})$ est positif. Par contre, le point x_{i1} est toujours inférieur à $\bar{x}_1$ dans ce sous-espace, si bien que le terme $(x_{i1} - \bar{x}_1)$ est négatif. En conséquence, le produit des deux termes $(x_{i1} - \bar{x}_1)(y_i - \bar{y})$ est négatif, car un nombre positif multiplié par un nombre négatif produit un nombre négatif. Il s'en suit que n'importe quel point dans le sous-espace # 2 contribue négativement à la covariance.

L'analyse pour les sous-espaces #3 et #4 se fait de manière similaire. La contribution des points au sous-espace #3 est positive car le terme $(x_{i1} - \bar{x}_1)(y_i - \bar{y})$ sera positif pour chaque observation. Similairement, la contribution des points du sous-espace #4 sera négative car le terme $(x_{i1} - \bar{x}_1)(y_i - \bar{y})$ sera négatif.

Quel sera le signe final de la covariance ? Tout dépend d'où se trouve la majo-

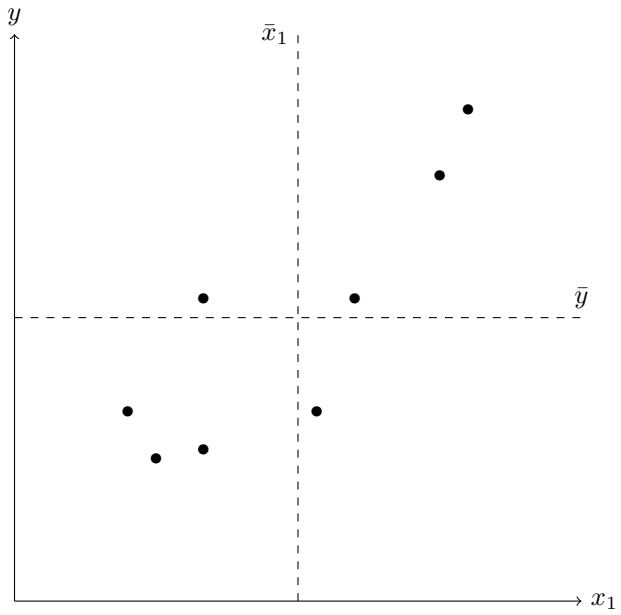

FIGURE 4 – La covariance de ces points est positive.

rité des points de l'échantillon. S'ils sont majoritairement dans les sous-espaces 1 et 3 (voir la figure 4), la covariance sera positive. S'ils sont majoritairement dans les sous-espaces 2 et 4 (voir la figure 5), la covariance sera négative. Finalement, s'ils sont dispersés plus ou moins uniformément dans les différents sous-espaces, la covariance sera nulle, ou proche de zéro (voir la figure 6).

En somme, une covariance positive traduit un mouvement conjoint qui est principalement positif alors qu'une covariance négative témoigne d'une relation principalement négative du mouvement joint des observations.

0.3.3 La variance : un cas particulier de la covariance

La covariance analyse le mouvement conjoint entre une variable x_1 et y. Or, quand on substitue la variable y par la variable x_1, la formule de covariance devient :

$$\begin{aligned} \text{cov}(x_1, x_1) &\overset{\text{def}}{=} \frac{1}{n-1}\sum_{i=1}^{n}(x_{i1} - \bar{x}_1)(x_{i1} - \bar{x}_1), \\ &= \frac{1}{n-1}\sum_{i=1}^{n}(x_{i1} - \bar{x}_1)^2 = \text{var}(x_1), \end{aligned} \tag{24}$$

soit la définition de la variance. En d'autres termes, la variance d'une variable statistique mesure à quel point une variable « bouge avec elle-même ». C'est une

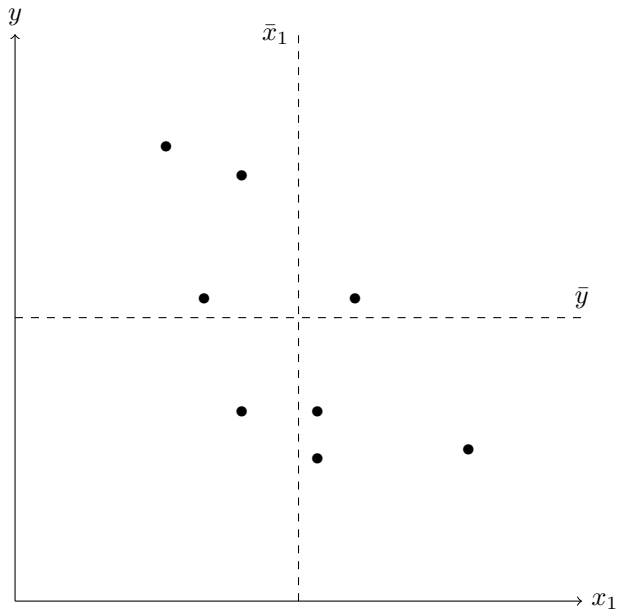

FIGURE 5 – La covariance de ces points est négative.

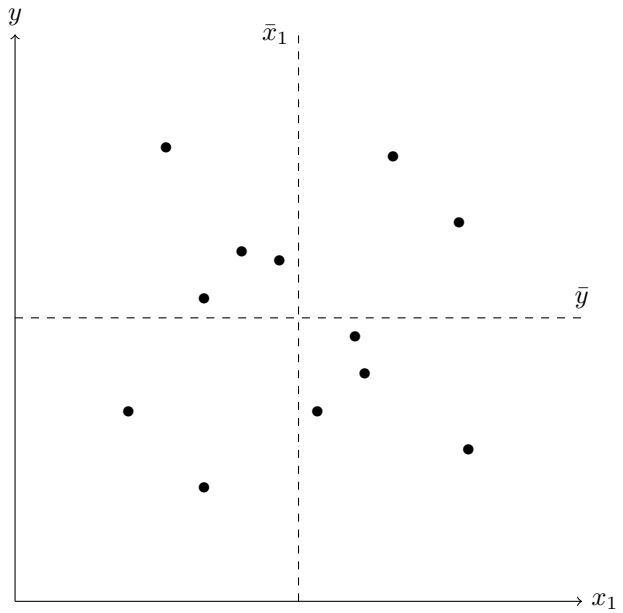

FIGURE 6 – La covariance de ces points est proche de zéro.

mesure de l'écart typique (du carré) d'une déviation par rapport à la moyenne. Intuitivement, la variance mesure à quel point les observations sont dispersées autour de la moyenne. Une variance élevée signifie que la dispersion est grande. Une faible variance signifie que les points sont très proches de la moyenne. Le cas extrême où la variance est nulle signifie que tout les points sont égaux à la moyenne.

0.3.4 Quelques propriétés de la covariance

Dans cette section, je vais établir un résultat associé à la covariance qui est excessivement utile pour les chapitres suivants. Soit x_i, y_i deux variables qui sont conjointement échantillonnées. Soit de plus la transformation $z_i = c + ax_i$ pour toutes les valeurs de i. Sachant la définition de la covariance entre x et y, on aimerait savoir quelle est la covariance entre y et z. Je montre ci-dessous que le résultat suivant est vrai :

$$\text{cov}(y, z) = a\text{cov}(y, x). \tag{25}$$

Pour le montrer, il suffit d'appliquer la définition de la covariance et de faire le développement algébrique associé :

$$\begin{aligned}
\text{cov}(y, z) &= \frac{1}{n-1} \sum_{i=1}^{n} (y_i - \bar{y}) (z_i - \bar{z}), \\
&= \frac{1}{n-1} \sum_{i=1}^{n} (y_i - \bar{y}) (c + ax_i - c - a\bar{x}), \\
&= \frac{1}{n-1} \sum_{i=1}^{n} (y_i - \bar{y}) (ax_i - a\bar{x}), \\
&= a\frac{1}{n-1} \sum_{i=1}^{n} (y_i - \bar{y}) (x_i - \bar{x}), \\
&= a\text{cov}(y, x).
\end{aligned}$$

En particulier, si $x_i = y_i$, on peut déduire que :

$$\text{cov}(y, z) = a\text{var}(x). \tag{26}$$

La deuxième propriété est aussi très utile, à savoir que :

$$\text{cov}(x, y) = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x}) (y_i - \bar{y}), \tag{27}$$

$$= \frac{1}{n-1} \sum_{i=1}^{n} (y_i - \bar{y}) (x_i - \bar{x}), \tag{28}$$

$$= \text{cov}(y, x) \tag{29}$$

TABLE 4 – Échantillon bidon pour fin d'exercices

i	x_{i1}	y_i	z_i
1	1	2	-0.5
2	2	4	- 2
3	3	6	-3.5
4	4	8	-5.5

. La covariance entre deux variables est la même, peu importe l'ordre dans lesquels on présente les arguments.

Finalement, la troisième propriété utile est le fait que la covariance entre une constante c et n'importe quelle variable x est toujours égal à zéro. Cela découle du fait que la moyenne d'une constante est égal à la constante elle-même :

$$\text{cov}(x, c) = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x}) \underbrace{(c - \bar{c})}_{=0}, \tag{30}$$

$$= 0. \tag{31}$$

0.3.5 Exercices

1. Calculer la moyenne des variables y et x_1 indiquées au tableau 1. Calculez de plus la covariance entre y et x_1, de même que la variance de x_1.

2. Soit l'échantillon présenté au tableau 4, calculez la covariance entre x_1 et y.

3. Toujours avec l'échantillon du tableau 4, calculer la covariance entre y et z.

4. Soit un système de points en x donné par l'échantillon $X = \{1, 2, 3, 4\}$ et soit le système de points en y généré par la relation $y_i = 3x_i + 1$. Servez-vous de l'intuition développée dans cette section pour indiquer si la covariance sera positive ou négative. Refaites la même analyse avec un système de points alternatifs $\tilde{y}_i = -3x + 1$. Tracez graphiquement.

5. Calculez les covariances associées à la question précédente.

0.3.6 Solutions

1. Les calculs sont illustrés ci-dessous :

$$
\begin{aligned}
\bar{x}_1 &= \frac{1}{3}\sum_{i=1}^{3} x_{i1} = \frac{1}{3}\left(24+26+30\right) = \frac{80}{3} \\
\bar{y} &= \frac{1}{3}\sum_{i=1}^{3} y_i = \frac{1}{3}\left(20+50+40\right) = \frac{110}{3} \\
\operatorname{cov}(x,y) &= \frac{1}{2}\sum_{i=1}^{3}\left(x_{1i}-\frac{80}{3}\right)\left(y_i-\frac{110}{3}\right) \\
&= \frac{1}{2}\left[\left(24-\frac{80}{3}\right)\left(20-\frac{110}{3}\right)+\left(26-\frac{80}{3}\right)\left(50-\frac{110}{3}\right)+\left(30-\frac{80}{3}\right)\left(40-\frac{110}{3}\right)\right] \\
&= \frac{1}{2}\left[\left(-\frac{8}{3}\right)\left(-\frac{50}{3}\right)+\left(-\frac{2}{3}\right)\left(\frac{40}{3}\right)+\left(\frac{10}{3}\right)\left(\frac{10}{3}\right)\right] \\
&= \frac{1}{18}\left[(-8)(-50)+(-2)(40)+(10)(10)\right] \\
&= \frac{1}{18}\left[400-80+100\right] \\
&= \frac{420}{18} = \frac{70}{3} \approx 23.333
\end{aligned}
$$

On peut remarquer que la covariance est positive, ce qui signale que lorsque la température augmente, le nombre de personnes à la piscine augmente. La variance est quant-à-elle donnée par :

$$
\begin{aligned}
\operatorname{var}(x_1) &= \frac{1}{2}\sum_{i=1}^{3}\left(x_{i1}-\bar{x}_1\right)^2 \\
&= \frac{1}{2}\left[\left(24-\frac{80}{3}\right)^2+\left(26-\frac{80}{3}\right)^2+\left(30-\frac{80}{3}\right)^2\right] \\
&= \frac{1}{2}\left[\left(-\frac{8}{3}\right)^2+\left(-\frac{2}{3}\right)^2+\left(\frac{10}{3}\right)^2\right] \\
&= \frac{1}{18}\left[64+4+100\right] \\
&= \frac{168}{18} = \frac{84}{9} \approx 9.333
\end{aligned}
$$

2. Les moyennes sont $\bar{x}_1 = \frac{5}{2}$ et $\bar{y} = 5$. La covariance est donnée par :

$$\begin{aligned}
\text{cov}(x_1, y) &= \frac{1}{3}\sum_{i=1}^{4}\left(x_{i1} - \frac{5}{2}\right)(y_i - 5) \\
&= \frac{1}{3}\left[\left(1 - \frac{5}{2}\right)(2-5) + \left(2 - \frac{5}{2}\right)(4-5) + \left(3 - \frac{5}{2}\right)(6-5) + \left(4 - \frac{5}{2}\right)(8-5)\right] \\
&= \frac{1}{3}\left[\left(-\frac{3}{2}\right)(-3) + \left(-\frac{1}{2}\right)(-1) + \left(\frac{1}{2}\right)(1) + \left(\frac{3}{2}\right)(3)\right] \\
&= \frac{1}{6}\left[(-3)(-3) + (-1)(-1) + (1)(1) + (3)(3)\right] \\
&= \frac{20}{6} = \frac{10}{3} \approx 3.333
\end{aligned}$$

3. La covariance est donnée par :

$$\text{cov}(y, z) = -5.5$$

4. La relation entre les points x_1 et y est positive dans le premier cas et négative dans le deuxième cas (voir la Figure 7). Puisque la covariance mesure si les points ont majoritairement une tendance croissante ou une tendance décroissante par rapport à la moyenne, la covariance sera positive dans le premier cas et négative dans le second.

5. $\text{cov}(x_1, y) = 5$ et $\text{cov}(x_1, \tilde{y}) = -5$. On peut montrer que pour une variable x_1 quelconque et une transformation sur cette variable $y = ax_1 + b$ on aura toujours $\text{cov}(x_1, y) = a\text{var}(x_1)$.[2] Ici, comme la variance de x_1 est égale à 5/3 et que la pente est soit 3, soit -3, le résultat s'en suit.

0.4 Régression à une variable

Les deux sections précédentes nous ont données des assises pour comprendre les fondements mathématiques d'une régression. Les formes quadratiques à plusieurs variables sont une généralisation des équations quadratiques que nous avons vues en révision. Leur sommet, ou leur point extrémum (minimum ou maximum, selon le signe) est donné par une généralisation de la formule $\frac{-b}{2a}$, ce qui donne un système d'équations qu'il faut résoudre. La covariance donne quant-à-elle une mesure du mouvement entre deux variables échantillonnées. Ces concepts seront utilisés ci-bas.

Dans cette section, nous revenons à la question énoncée en introduction, soit scomment déterminer les coefficients β_0, β_1 qui nous permettent de bien représenter les données observées liant la température au nombre des personnes à la piscine ? Autrement dit, quels sont les coefficients de l'équation 1 ?

Pour comprendre quel critère est employé pour déterminer les coefficients de régression, je commence par tracer les points de l'échantillon reporté au tableau

2. Nous allons discuter formellement cette propriété plus tard dans le cours.

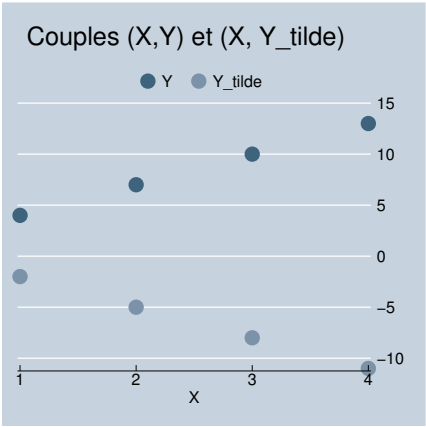

FIGURE 7 – Couples (x_i, y_i) et $(x_i, \tilde{y}_i)$

1 à la Figure 8 et j'y trace une droite « quelconque » comme possible candidate de droite représentant l'échantillon. Sur cette droite, j'ai présenté chaque valeur prédite, soit les prédictions du nombre de personnes pour les températures reportées. Ces points sont représentés par une croix. J'ai aussi illustré chaque variable d'erreur de prévision (e_1, e_2, e_3) entre cette droite hypothétique et la valeur observée. Ces variables d'erreur de prévision sont présentées par les lignes pointillées. Notez qu'en traçant une droite de prévision, on définit nécessairement la valeur des écarts de prédiction.

À la Figure 9, je présente une autre droite candidate qui pourrait servir de prévision. Une inspection visuelle de cette droite devrait intuitivement mener au rejet de cette droite come « bonne » candidate pour représenter les points échantillonnés. En effet, la relation entre la température et le nombre de personne est positive alors que la pente de cette seconde droite candidate est négative. Clairement, la droite de la Figure 8 représente mieux les données ! Mais qu'entend-on exactement par « mieux » ? Quel critère formel emploie-t-on ?

Si on prend une règle, on peut constater que la somme de la longueur des erreurs de prédiction à la Figure 9 est beaucoup plus grande que celle à la Figure 8. Un critère formel qui est attrayant est donc le suivant : on aimerait que l'ensemble des erreurs de prédiction, soit la somme des écarts entre chaque variable observée et sa valeur prédite associée, soit la plus petite possible. Une prédiction qui, justement, *minimise* les erreurs de prédiction est clairement quelque chose de désirable !

Pour mesurer adéquatement la notion de distance entre les prédictions et les valeurs réelles, indépendamment que cette distance soit positive (prédiction inférieure à la valeur réelle) ou négative (prédiction supérieure à la valeur réelle), on emploie le *carré* des erreurs de prédiction. Pour un échantillon de taille n, une droite de prédiction génèrera n erreurs de prédiction $e_1, e_2, \ldots, e_n$. On cherchera donc à choisir les coefficients β_0 et β_1 qui minimisent la somme du carré des

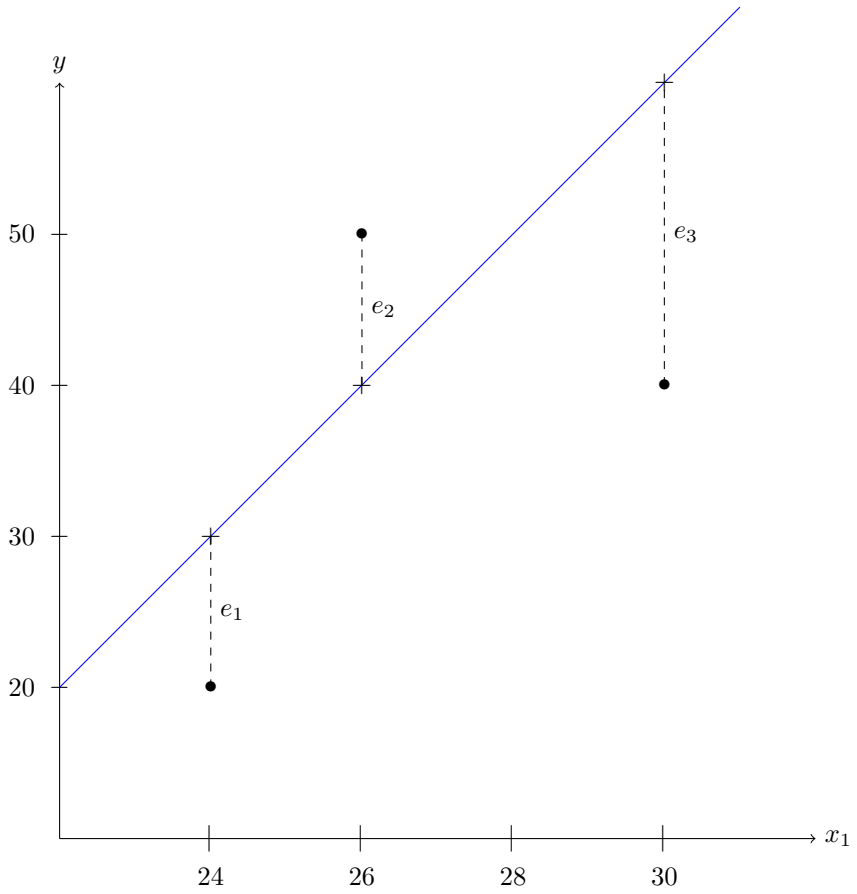

FIGURE 8 – Points du tableau 1, droite quelconque de prévision et les écarts de prédiction associés.

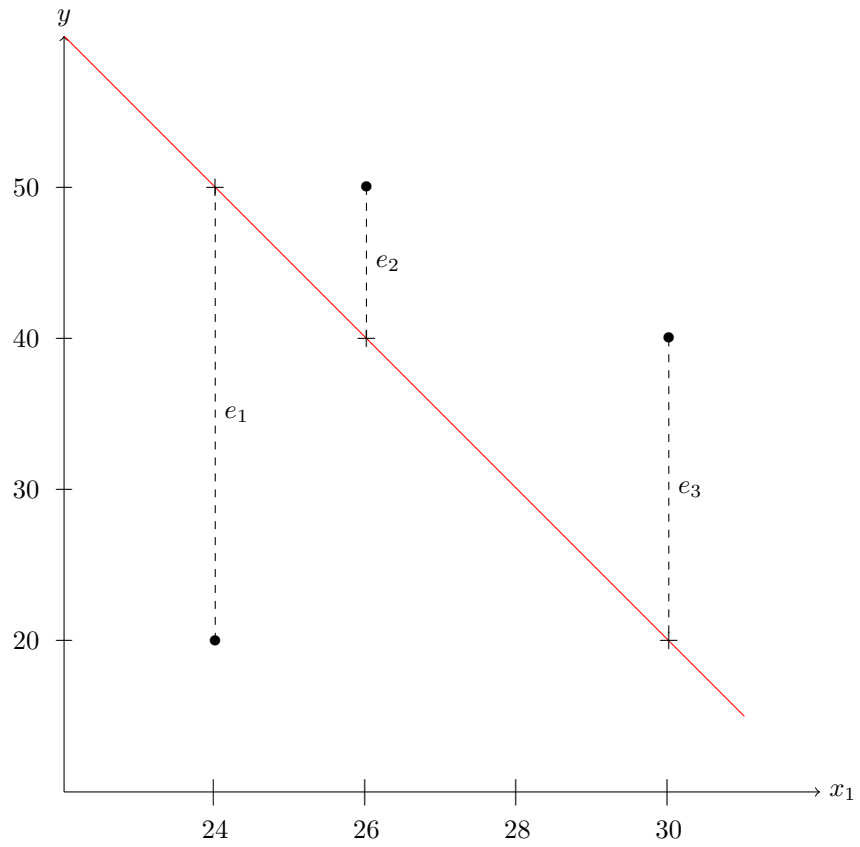

FIGURE 9 – Points du tableau 1, une autre droite quelconque de prévision et les écarts de prédiction associés.

termes d'erreur $\left(\sum_i e_i^2\right)$. Mathématiquement, on écrit :

$$\min SSR = \min \sum_{i=1}^{n} e_i^2. \tag{32}$$

Le terme SSR vient de l'anglais et désigne *sum of square residuals* (somme du carré des résidus, ou somme des erreurs de prédiction). Par l'équation 1, on peut écrire l'erreur de prédiction e_i comme l'écart entre la valeur observée et la valeur prédite : $e_i = y_i - \beta_0 - \beta_1 x_{i1}$. L'expression $y_i - \beta_0 - \beta_1 x_{i1}$ est la représentation mathématique de la ligne pointillée associée à l'observation i dans la Figure 8. On peut donc remplacer le terme e_i^2 par cette expression dans la fonction SSR :

$$SSR = \sum_{i=1}^{n} \left(y_i - \beta_0 - \beta_1 x_{i1}\right)^2. \tag{33}$$

Choisir β_0 et β_1 de manière à minimiser SSR revient donc à choisir la droite qui minimise la longueur des traits pointillés dans la 8, soit la somme (du carré) des erreurs de prédiction.

Comme les prochaines étapes contiennent un peu plus de développements mathématiques, des explications préalables sont utiles. L'idée des prochains paragraphes est de montrer deux choses. D'une part, l'expression SSR est une forme quadratique à deux variables (soit β_0 et β_1), si bien que la minimiser consiste à trouver son sommet, ce qu'on a appris à faire à la section 0.2. D'autre part, les formules déterminant la valeur de β_0 et β_1 s'expriment en fonction de moyennes, de variances et de covariances, concepts que nous avons couvert à la section 0.3. Bref, les développements qui viennent ci-dessous permettent de trouver les expressions pour β_0^* et β_1^*.

Commençons par illustrer que l'expression pour SSR est bien une forme quadratique. Pour ce faire, il faut développer l'expression sur laquelle s'applique la sommation :

$$\begin{aligned} SSR &= \sum_{i=1}^{n} \left(y_i - \beta_0 - \beta_1 x_{i1}\right)^2 \\ &= \sum_{i=1}^{n} \left(y_i^2 - 2\beta_0 y_i - 2\beta_1 x_{i1} y_i + 2\beta_0\beta_1 x_{i1} + \beta_0^2 + \beta_1^2 x_{i1}^2\right). \end{aligned} \tag{34}$$

Ensuite, il faut distribuer la sommation :

$$\begin{aligned} SSR &= \sum_{i=1}^{n} y_i^2 - 2\beta_0 \sum_{i=1}^{n} y_i - 2\beta_1 \sum_{i=1}^{n} x_{i1} y_i + 2\beta_0\beta_1 \sum_{i=1}^{n} x_{i1} + \underbrace{\sum_{i=1}^{n} \beta_0^2}_{=n\beta_0^2} + \beta_1^2 \sum_{i=1}^{n} x_{i1}^2, \\ &= \beta_1^2 \left(\sum_{i=1}^{n} x_{i1}^2\right) + \sum_{i=1}^{n} y_i^2 - \beta_0 \left(2\sum_{i=1}^{n} y_i\right) - \beta_1 \left(2\sum_{i=1}^{n} x_{i1} y_i\right) + \beta_0\beta_1 \left(2\sum_{i=1}^{n} x_{i1}\right) + n\beta_0^2. \end{aligned} \tag{35}$$

Rappelez-vous que chacune des sommations dans cette expression correspond à un nombre qui est implicitement défini par l'échantillon du problème. Pour l'illustrer, je remplace ci-dessous chacun des termes de sommation par leur valeur tirée du tableau 4 :

$$SSR = \underbrace{2152}_{\sum_i x_{i1}^2}\beta_1^2 + \underbrace{4500}_{\sum_i y_i^2} - \underbrace{220}_{2\sum_i y_i}\beta_0 - \underbrace{5960}_{2\sum_i x_{i1}y_i}\beta_1 + \underbrace{160}_{2\sum_i x_{i1}}\beta_0\beta_1 + \underbrace{3}_{n}\beta_0^2, \tag{36}$$

Puisqu'on connaît les termes numériques associés à cet exemple, une lectrice pourrait se demander pourquoi on ne reste pas avec ces nombres pour résoudre le problème tout en simplifiant l'exposé. Après-tout, il est plus facile de lire l'équation 36 que l'équation 35. L'idée est qu'en gardant la symbolique mathématique abstraite (l'équation 35), on peut trouver une réponse générale, une réponse qui s'applique à *toutes* les situations, et non juste celui associé à l'introduction. En somme, faire un effort de développement et de compréhension sur cette équation abstraite permet de généraliser. L'abstraction est plus difficile, mais offre une réponse « universelle ».

En réorganisant les termes de l'équation 35, on peut montrer que sa structure correspond à une équation quadratique en β_1 :

$$SSR = \beta_1^2 \underbrace{\left(\sum_{i=1}^{n} x_{i1}^2\right)}_{a} - \beta_1 \underbrace{\left(\left(2\sum_{i=1}^{n} x_{i1}y_i\right) - \beta_0\left(2\sum_{i=1}^{n} x_{i1}\right)\right)}_{b} + \underbrace{\left(\sum_{i=1}^{n} y_i^2 - \beta_0\left(2\sum_{i=1}^{n} y_i\right) + n\beta_0^2\right)}_{c} \tag{37}$$

On sait que la valeur minimale de cette expression se trouve à son sommet, qui est donné à $-\frac{b}{2a}$:

$$\begin{aligned}\beta_1^* &= \frac{2\sum_{i=1}^{n} x_{i1}y_i - \beta_0^* 2\sum_{i=1}^{n} x_{i1}}{2\sum_{i=1}^{n} x_{i1}^2}, \\ &= \frac{\sum_{i=1}^{n} x_{i1}y_i - \beta_0^* \sum_{i=1}^{n} x_{i1}}{\sum_{i=1}^{n} x_{i1}^2}\end{aligned} \tag{38}$$

Maintenant, je reprends l'expression de SSR et la développe pour expliciter sa structure quadratique en β_0 :

$$SSR = \underbrace{n}_{a}\beta_0^2 - \beta_0 \underbrace{\left(\left(2\sum_{i=1}^{n} y_i\right) - \beta_1\left(2\sum_{i=1}^{n} x_{i1}\right)\right)}_{b} + \underbrace{\left(\beta_1^2\left(\sum_{i=1}^{n} x_{i1}^2\right) + \sum_{i=1}^{n} y_i^2 - \beta_1\left(2\sum_{i=1}^{n} x_{i1}y_i\right)\right)}_{c}. \tag{39}$$

Le minimum en β_0 est aussi donné par l'application de la formule $-\frac{b}{2a}$ à l'ex-

pression :

$$\begin{aligned}
\beta_0^* &= \frac{2\sum_{i=1}^n y_i - \beta_1^* 2\sum_{i=1}^n x_{i1}}{2n}, \\
&= \frac{\sum_{i=1}^n y_i - \beta_1^* \sum_{i=1}^n x_{i1}}{n}, \\
&= \underbrace{\frac{1}{n}\sum_{i=1}^n y_i}_{=\bar{y}} - \beta_1^* \underbrace{\frac{1}{n}\sum_{i=1}^n x_{i1}}_{=\bar{x}_1}, \\
\Rightarrow \beta_0^* &= \bar{y} - \beta_1^* \bar{x}_1. \qquad (40)
\end{aligned}$$

Cette dernière équation est relativement simple : l'ordonnée à l'origine qui minimise les erreurs de prédiction est donnée par la différence entre la moyenne du nombre de personnes observées et la moyenne de la température observée, multipliée par le coefficient β_1^*.

Un problème demeure, à savoir qu'on n'a toujours pas de solution pour le coefficient β_1^*. Son expression dépend toujours de β_0^*. Il faut donc substituer l'équation de β_0^* dans l'équation 38 et trouver une forme explicite pour β_1^* :

$$\begin{aligned}
\beta_1^* &= \frac{\sum_{i=1}^n x_{i1} y_i - \beta_0^* \sum_{i=1}^n x_{i1}}{\sum_{i=1}^n x_{i1}^2}, \\
&= \frac{\sum_{i=1}^n x_{i1} y_i - (\bar{y} - \beta_1^* \bar{x}_1)\sum_{i=1}^n x_{i1}}{\sum_{i=1}^n x_{i1}^2}, \\
&\qquad \text{(substitution)} \\
&= \frac{\sum_{i=1}^n x_{i1} y_i - \bar{y}\sum_{i=1}^n x_{i1} + \beta_1^* \bar{x}_1 \sum_{i=1}^n x_{i1}}{\sum_{i=1}^n x_{i1}^2}, \\
&\qquad \text{(distribution)} \\
&= \frac{\sum_{i=1}^n x_{i1}(y_i - \bar{y}) + \beta_1^* \bar{x}_1 \sum_{i=1}^n x_{i1}}{\sum_{i=1}^n x_{i1}^2}, \\
&\qquad \text{(regroupement en } x_{i1}\text{)} \\
\Rightarrow \sum_{i=1}^n x_{i1}^2 \beta_1^* &= \sum_{i=1}^n x_{i1}(y_i - \bar{y}) + \beta_1^* \bar{x}_1 \sum_{i=1}^n x_{i1} \\
&\qquad \text{(multiplication du dénominateur)} \\
\Rightarrow \sum_{i=1}^n x_{i1}^2 \beta_1^* - \beta_1^* \bar{x}_1 \sum_{i=1}^n x_{i1} &= \sum_{i=1}^n x_{i1}(y_i - \bar{y}) \qquad \text{(termes en } \beta_1^* \text{ à gauche)} \\
\Rightarrow \beta_1^* \left(\sum_{i=1}^n x_{i1}^2 - \bar{x}_1 \sum_{i=1}^n x_{i1} \right) &= \sum_{i=1}^n x_{i1}(y_i - \bar{y}) \\
&\qquad \text{(factorisation des termes en } \beta_1\text{)} \\
\Rightarrow \beta_1^* \sum_{i=1}^n (x_{i1} - \bar{x}_1) x_{i1} &= \sum_{i=1}^n x_{i1}(y_i - \bar{y}). \qquad (41)
\end{aligned}$$

Pour finaliser le développement de la formule, il faut utiliser le truc développé à l'équation 22, à savoir que :

$$0 = \bar{y} \sum_{i=1}^{n} (x_i - \bar{x}).$$

Comme ce terme est égal à zéro, on peut l'ajouter d'un côté ou de l'autre de l'équation sans changer l'égalité. Le fait d'ajouter ce terme permettra cependant de simplifier l'expression finale, ce qui facilitera l'interprétation. Chacun de ces ajouts est souligné par un signe $= 0$ pour faciliter le suivi. Le reste des étapes sont des manipulations algébriques réorganisant les équations :

$$\beta_1^* \sum_{i=1}^{n} (x_{i1} - \bar{x}_1) x_{i1} = \sum_{i=1}^{n} x_{i1}(y_i - \bar{y}) - \underbrace{\bar{x}_1 \sum_{n=1}^{n} (y_i - \bar{y})}_{=0},$$

(ajout d'un terme nul)

$$\Rightarrow \beta_1^* \sum_{i=1}^{n} (x_{i1} - \bar{x}_1) x_{i1} = \sum_{i=1}^{n} (x_{i1} - \bar{x}_1)(y_i - \bar{y}),$$

(regroupement des termes)

$$\Rightarrow \beta_1^* \frac{1}{n-1} \sum_{i=1}^{n} (x_{i1} - \bar{x}_1) x_{i1} = \underbrace{\frac{1}{n-1} \sum_{i=1}^{n} (x_{i1} - \bar{x}_1)(y_i - \bar{y})}_{=\mathrm{cov}(x_1, y)}.$$

(multiplication par $\frac{1}{n-1}$)

On devrait remarquer que le terme de droite de la dernière expression est la définition de la covariance entre x_1 et y. C'est une expression qu'on connaît et qu'on est capable d'interpréter. Il reste à simplifier le terme de gauche, ce qu'on va aussi faire en ajoutant un terme égal à zéro :

$$\Rightarrow \beta_1^* \frac{1}{n-1} \sum_{i=1}^{n} (x_{i1} - \bar{x}_1) x_{i1} - \underbrace{\frac{1}{n-1} \bar{x}_1 \sum_{i=1}^{n} (x_{i1} - \bar{x}_1)}_{=0} = \mathrm{cov}(x_1, y),$$

(ajout d'un terme nul)

$$\Rightarrow \beta_1^* \underbrace{\frac{1}{n-1} \sum_{i=1}^{n} (x_{i1} - \bar{x}_1)^2}_{=\mathrm{var}(x_1)} = \mathrm{cov}(x_1, y), r$$

(regroupement des termes)

$$\Rightarrow \beta_1^* \mathrm{var}(x_1) = \mathrm{cov}(x_1, y),$$

(simplification)

$$\Rightarrow \beta_1^* = \frac{\mathrm{cov}(x_1, y)}{\mathrm{var}(x_1)}.$$

(résolution)

L'utilisation du truc développé à l'équation 22 permet donc d'exprimer la formule de β_1^* en termes de concepts plus intelligibles, soit ceux de la covariance entre x_1 et y et de la variance de x_1.

Toute la démarche des pages précédentes nous aura permis de trouver l'expression pour les deux coefficients β_0^* et β_1^* qui minimisent le carré des erreurs de prédictions. Ces expressions sont données par :

$$\beta_1^* = \frac{\text{cov}(x_1, y)}{\text{var}(x_1)}, \tag{42}$$

$$\beta_0^* = \bar{y} - \beta_1^* \bar{x}_1. \tag{43}$$

Remarquez que les observations n'apparaissent pas explicitement dans ces formules : les deux expressions sont valides peu importe la taille de l'échantillon (n) car on sait calculer les moyennes, covariances et variances pour toute taille d'échantillon. En conséquence, les formules développées sont générales.

Le coefficient pour la pente, β_1^*, dépend au numérateur de la covariance entre x_1 et y et de la variance de x_1 au dénominateur. Comme la variance est un nombre qui est toujours positif, il s'en suit que le signe de la pente héritera du signe de la covariance. Or, on sait que le signe de la covariance dépend de la disposition des points observés par rapport à la moyenne. S'ils sont majoritairement disposés comme une droite à pente positive (sous-espaces # 1 et # 3), la covariance sera positive. Conséquemment, la pente estimée sera aussi positive. Inversement, si les points observés sont majoritairement disposés comme une droite à pente négative (sous-espaces #2 et # 4), la covariance sera négative. En conséquence, la pente sera aussi négative. Bref, le signe de la pente des prévisions est déterminée par l'orientation générale des points observés, ce que traduit l'expression de la covariance.

La magnitude de la pente dépend quant-à-elle de deux facteurs, soit l'ampleur de la covariance entre x_1 et y et l'ampleur de la variance de x_1. Plus la covariance est grande, plus la magnitude de la pente sera grande. Plus la variance sur x_1 est petite, plus le coefficient sera grand.

0.4.1 Trois propriétés importantes de la droite de régression estimée

La droite de régression estimée a trois propriétés importantes. Dans un premier temps, elle passe nécessairement par la moyenne de l'échantillon, c'est-à-dire que si on cherche la prédiction au point $\bar{x}_1$, on va obtenir :

$$\begin{aligned} \beta_0^* + \beta_1^* \bar{x}_1 &= (\bar{y} - \beta_1^* \bar{x}_1) + \beta_1^* \bar{x}_1 \\ &= \bar{y}. \end{aligned} \tag{44}$$

Bref, la droite de régression passe par le point $(\bar{x}_1, \bar{y})$.

Dans un deuxième temps, la moyenne des erreurs d'estimation est nulle $\left(\frac{1}{n}\sum_{i=1}^{n} e_i = 0\right)$. Pour s'en rendre compte, il faut faire la sommation de l'équa-

tion estimée de chaque côté :

$$\begin{aligned}
y_i &= \beta_0^* + \beta_1^* x_{i1} + e_i, && \text{(equation originale)}\\
\Rightarrow \sum_{i=1}^{n} y_i &= \sum_{i=1}^{n} \beta_0^* + \beta_1^* \sum_{i=1}^{n} x_{i1} + \sum_{i=1}^{n} e_i && \text{(sommation de chaque côté)}\\
\Rightarrow \frac{1}{n}\sum_{i=1}^{n} y_i &= \frac{1}{n}\sum_{i=1}^{n} \beta_0^* + \beta_1^* \frac{1}{n}\sum_{i=1}^{n} x_{i1} + \frac{1}{n}\sum_{i=1}^{n} e_i && \text{(multiplication par 1/n)}\\
\bar{y} &= \beta_0^* + \beta_1^* \bar{x}_1 + \frac{1}{n}\sum_{i=1}^{n} e_i && \text{(par définition des moyennes)}
\end{aligned}$$

On sait, par la dernier résultat discuté, que $\bar{y} = \beta_0^* + \beta_1^* \bar{x}_1$, si bien qu'on peut éliminer ces termes de l'expression, ce qui implique :

$$0 = \frac{1}{n}\sum_{i=1}^{n} e_i. \tag{45}$$

soit précisément le résultat recherché. En moyenne, la droite d'estimation développée ne fait pas d'erreur de prévision.

Finalement, la covariance entre les termes d'erreur estimés et la prévision de la droite est égale à zéro. On peut le voir en développant le calcul de covariance associé, où l'usage des propriétés de la covariance est très utile :

$$\begin{aligned}
\mathrm{cov}(\beta_0^* + \beta_1^* x_1, e) &= \mathrm{cov}(\beta_1^* x_1, e), && \text{(Propriété de la covariance)}\\
&= \beta_1^* \mathrm{cov}(x_1, e), && \text{(Propriété de la covariance)}\\
&= \beta_1^* \mathrm{cov}(x_1, y - \beta_0^* - \beta_1^* x_1), && \text{(définition des résidus)}\\
&= \beta_1^* \mathrm{cov}(x_1, y - \beta_1^* x_1), && \text{(Propriété de la covariance)}\\
&= \beta_1^* \mathrm{cov}(x_1, y) - \beta_1^* \mathrm{cov}(x_1, \beta_1^* x_1), && \text{(Propriété de la covariance)}\\
&= \beta_1^* \mathrm{cov}(x_1, y) - (\beta_1^*)^2 \mathrm{var}(x_1), && \text{(Propriété de la covariance)}\\
&= \frac{\mathrm{cov}(x_1, y)}{\mathrm{var}(x_1)} \mathrm{cov}(x_1, y) - \left(\frac{\mathrm{cov}(x_1, y)}{\mathrm{var}(x_1)}\right)^2 \mathrm{var}(x_1), && \text{(Valeur de } \beta_1)\\
&= 0. && (46)
\end{aligned}$$

Ces trois propriétés reviendront systématiquement dans n'importe quelle analyse de régression.

0.4.2 Application aux données du tableau 1

Ces formules étant maintenant développées, on peut trouver les coefficients qui représentent le mieux les données observées par l'équipe de planification opérationnelle. On peut trouver une droite qui génère la prédiction minimisant les erreurs de prévision.

Dans les exercices de la section 0.3, on a calculé les valeurs de $\text{cov}(x_1, y), \text{var}(x_1), \bar{x}_1$ et $\bar{y}$:

$$\bar{x}_1 = \frac{80}{3} \qquad \bar{y} = \frac{110}{3}$$
$$\text{cov}(x_1, y) = \frac{70}{3} \qquad \text{var}(x_1) = \frac{84}{9}.$$

On peut donc s'en servir pour calculer les coefficients estimés qui minimisent le carré des termes d'erreur :

$$\begin{aligned}
\beta_1^* &= \frac{\text{cov}(x_1, y)}{\text{var}(x_1)}, \\
&= \frac{70}{3}\frac{9}{84}, \\
&= \frac{5}{2} = 2.5 \qquad (47) \\
\beta_0^* &= \bar{y} - \beta_1^* \bar{x}_1, \\
&= \frac{110}{3} - 2.5\frac{80}{3} \\
&= -\frac{90}{3} = -30. \qquad (48)
\end{aligned}$$

Bref, la prévision du nombre de personnes en fonction de la température qui minimise les erreurs de prévision est donnée par la relation :

$$y_i = -30 + 2.5x_{i1}. \qquad (49)$$

Cette droite est présentée graphiquement à la Figure 10.

Avec cette équation, les erreurs de prédiction sont respectivement $e_1^* = -10, e_2^* = 15, e_3^* = -5$, ce qui génère une somme du carré des erreurs de prédiction de :

$$\begin{aligned}
SSR &= (-10)^2 + 15^2 + (-5)^2, \\
&= 350. \qquad (50)
\end{aligned}$$

Par construction de l'analyse de régression qu'on vient de faire, on sait que cette valeur est minimale car on emploie la relation qui fait le moins d'erreurs. On remarque aussi que la somme des termes d'erreurs, ou la moyenne des termes d'erreurs, est nulle :

$$\frac{1}{3}\sum_{i=1}^{3} e_i = \frac{1}{3}(-10 + 15 - 5) = 0. \qquad (51)$$

Prévision : une introduction

On peut se servir de l'équation trouvée pour faire une prévision du nombre de personnes qui seraient à la piscine si la température était de 28 degrés Celsius.

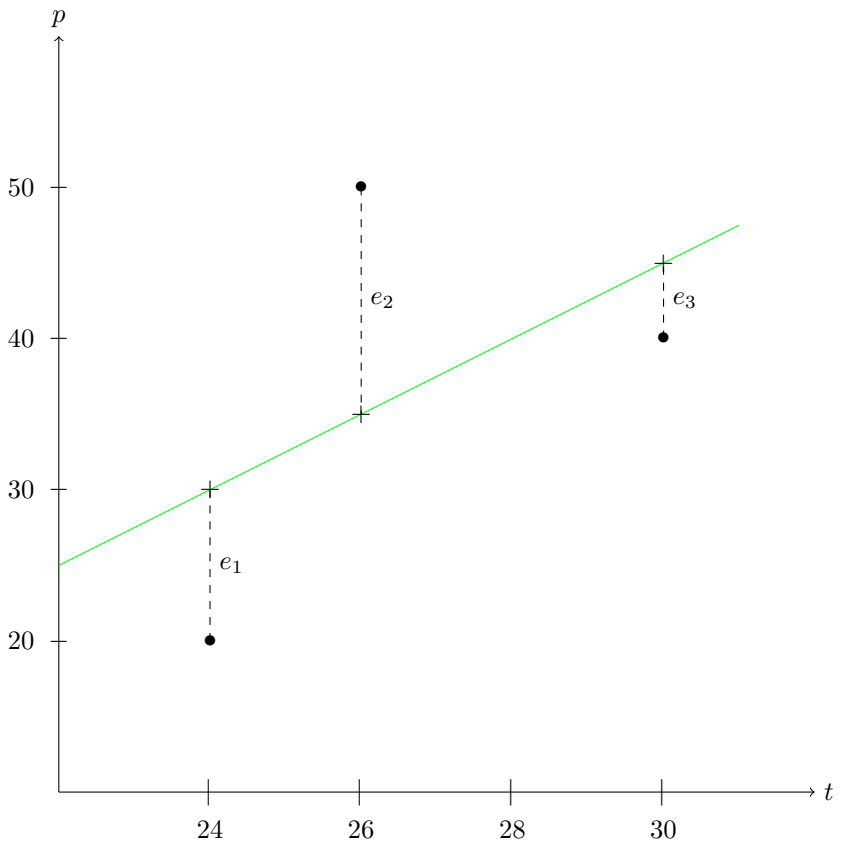

FIGURE 10 – Points du tableau 1 et droite qui minimise les écarts de prédiction.

Il suffit de se servir de l'équation estimée en utilisant la valeur $x_{i1} = 28$:

$$\begin{aligned} y_i &= -30 + 2.5 \cdot 28, \\ &= 40. \end{aligned} \tag{52}$$

En somme, notre modèle prévoît que s'il fait 28 degrés Celsius, on peut s'attendre à 40 personnes à la piscine. Cette prévision nous donne la meilleure valeur basée sur notre modèle d'estimation. Il reste cependant deux questions à résoudre. D'une part, à quel point la prévision qu'on vient de faire est fiable, c'est-à-dire, quelle est l'incertitude rattachée à cette prévision ? D'autre part, comment déterminer si le modèle que nous avons postulé ne peut pas être substitué par un *meilleur* modèle de prévision ? Pour répondre à ces questions, il faut développer la dimension statistique aux régressions, ce que nous ferons ultérieurement.

0.4.3 Exercices

1. Reprenez les données du tableau 4 est estimez le modèle de régression suivant :

$$y_i = \beta_0 + \beta_1 x_i + e_i. \tag{53}$$

   Expliquez pourquoi cette estimation génère un terme d'erreur qui est nul pour chaque observation.

2. Toujours avec les données du tableau 4, estimez le modèle de régression :

$$z_i = \beta_0 + \beta_1 x_i + e_i. \tag{54}$$

3. Soit le modèle de prévisions $y_i = \beta_0 + e_i$, un qui cherche à estimer le nombre de personnes à la piscine par une constante. Dérivez la formule pour β_0^* qui minimise le carré des termes d'erreurs.

0.4.4 Solutions

1. La covariance entre x et y est de $\frac{10}{3}$. La variance de x est de $\frac{5}{3}$. En conséquence, le coefficient d'estimation β_1^* est :

$$\beta_1^* = \frac{10/3}{5/3} = 2. \tag{55}$$

   Le coefficient β_0^* est quant-à-lui donné par $\bar{y} - 2\bar{x}$ et comme $\bar{y} = 5$ et $\bar{x} = \frac{5}{2}$, on peut en conclure que $\beta_0^* = 0$.

   Les termes d'erreurs de cette régression sont tous nuls ($e_i = 0, i = 1...4$) parce que la relation entre les variables x_i et y_i est parfaitement linéaire. La droite estimée épouse parfaitement les données !

2. La covariance entre x et z est $-\frac{11}{4}$ et la variance de x est toujours de $\frac{5}{3}$. En conséquence, le coefficient estimé est :

$$\beta_1^* = -\frac{33}{20} = -1.65 \tag{56}$$

   Le coefficient estimé β_0^* dépend quant-à-lui de la moyenne en x et en z. La moyenne en x est toujours de $\frac{5}{2}$ et la moyenne en z est $-\frac{23}{8}$. L'estimateur est donc donné par :

$$\beta_0^* = -\frac{23}{8} + 1.65\frac{5}{2} = \frac{5}{4}. \tag{57}$$

   En somme, la droite recherchée est donnée par :

$$z_i = \frac{5}{4} - \frac{33}{20}x_i + e_i. \tag{58}$$

3. La somme du carré des termes d'erreurs est donnée par :

$$\begin{aligned}\sum_i e_i^2 &= \sum_i (y_i - \beta_0)^2, \\ &= \underbrace{\sum_i y_i^2}_{=c} - \beta_0 \underbrace{2\sum_i p_i}_{=b} + \underbrace{n}_{=a} \beta_0^2.\end{aligned}$$

C'est une forme quadratique univariée en β_0. Son minimum est à $-\frac{b}{2a}$, ce qui nous donne :

$$\begin{aligned}\beta_0^* &= \frac{2\sum_i y_i}{2n}, \\ &= \frac{\sum_i y_i}{n} = \bar{y}.\end{aligned} \tag{59}$$

La constante qui minimise le carré des erreurs de prédiction est donné par la moyenne des observations ($\bar{y}$).

0.5 Régression à plusieurs variables

Armé du modèle de prévision identifié dans les sections précédentes, l'équipe de planification des opérations présente ses résultats à la direction de la ville. Pendant la rencontre, il ne faut pas beaucoup de temps pour qu'un sauveteur de la ville voit immédiatement les limitations du modèle : « mais qu'en est-il des fins de semaines ?», questionne-t-il, « il y a toujours plus de monde quand c'est la fin de la semaine ! »

De toute évidence, le modèle développé ne prends en considération que la température de la journée, sans tenir compte du jour de la semaine. On aimerait aussi tenir compte du fait qu'on est un jour de semaine ou un jour de fin de semaine. En somme, on aimerait que le modèle de prévision dépende de plusieurs variables dépendantes.

L'équipe de planification retourne donc à la planche à dessin. Ils se souviennent dans un premier temps que les premières observations furent effectuées pendant les jours de semaine. Ils doivent donc prendre de nouvelles mesures la fin de semaine pour évaluer l'impact de la fin de semaine sur le nombre de personnes à la piscine. Soit x_2 la variable qui désigne si les observations sont prises pendant la fin de semaine ou non. Cette variable sera égale à 1 si l'observation est prise pendant la fin de semaine et égale à zéro si elle est prise pendant la semaine. Ce type de variable est ce qu'on appelle une **variable binaire**, ou une **variable dichotomique**, car la variable ne peut prendre que deux valeurs. Les données complètes recueillies sont présentées au tableau 5.

Ce nouvel échantillon a maintenant 6 observations ($n = 6$). Par exemple, la sixième observation rapporte qu'il y avait 65 personnes à la piscine ($y_i = 65$) alors qu'il faisait 28 degrés Celsius ($x_{i1} = 28$), une journée de fin de semaine ($x_{i2} = 1$).

TABLE 5 – Nouvel échantillon recueilli par l'équipe de planification

Observation i	Température x_{i1}	Personnes y_i	Fin de semaine x_{i2}
1	24	20	0
2	26	50	0
3	30	40	0
4	30	60	1
5	24	45	1
6	28	65	1

On aimerait alors développer un nouveau modèle de régression linéaire qui permet d'expliquer le nombre de personnes à la piscine en fonction des deux variables :

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + e_i. \tag{60}$$

Ce nouveau modèle de régression est similaire au modèle précédent, à l'exception du fait qu'il ajoute maintenant une nouvelle variable dépendante, soit la variable « x_{i2} » désignant si l'observation est mesurée la fin de semaine ou non. Puisqu'il y a une nouvelle variable, il y a un nouveau coefficient à estimer, β_2, qui représente l'impact de la variable « fin de semaine » sur le nombre de personnes à la piscine.

Qu'il serait confortant de savoir que l'estimation de ce coefficient pourrait se faire en suivant les formules développées à la section précédente ! Malheureusement, ce n'est pas le cas. Nous allons voir que le fait d'ajouter une variable fait en sorte que les formules développées pour les coefficients β_0 et β_1 ne tiennent plus quand on ajoute une (ou plusieurs autres) variable. Il faut développer de nouvelles formules et refaire l'analyse.

On peut alors se demander à quoi aura servi la section précédente si les formules développées ne tiennent que lorsqu'il y a une seule variable. Pourquoi tant de mal si tout est à refaire ? La réponse se décline en deux temps.

D'abord, la *recette* qui mène au développement des estimateurs, lorsqu'il y a plusieurs variables, est la même que lorsqu'il y en a qu'une seule. On cherche encore à choisir les coefficients β_0, β_1 et β_2 qui minimisent la somme du carré des erreurs de prédiction. Le SSR est encore une forme quadratique, cette fois à trois variables, qu'on peut minimiser en identifiant son sommet. Finalement, les formules estimées sont aussi la solution au système d'équation identifiant le sommet du SSR. Ce système a cette fois trois équations et trois inconnues. En plus d'une recette similaire, les trucs algébriques développés dans la section précédentes s'appliquent encore pour cette section. En somme, c'est la réponse qui change et non la recette. Si vous avez compris la recette, vous comprendrez cette section relativement rapidement.

Ensuite, le développement des formules associées au cas avec une seule va-

riable aura permis de développer de l'intuition. On sait que le coefficient β_1 dépend de la covariance entre la variable dépendante et la variable indépendante. Nous allons voir que la nouvelle formule sera légèrement différente, mais le fait d'avoir vu la première version facilitera son interprétation.

0.5.1 Coefficients de régression et interprétation

Pour un modèle à deux variables indépendantes, l'équation du carré des erreurs de prédiction est donné par :

$$SSR = \sum_i e_i^2 = \sum_i (y_i - \beta_0 - \beta_1 x_{i1} - \beta_2 x_{i2})^2. \tag{61}$$

En développant l'argument de la sommation et en le distribuant, on trouve :

$$\begin{aligned} SSR &= \sum_i (y_i^2 - 2\beta_0 y_i - 2\beta_1 y_i x_{i1} - 2\beta_2 y_i x_{i2} + 2\beta_1\beta_0 x_{i1} + 2\beta_0\beta_2 x_{i2} + \dots \\ &\qquad \cdots + 2\beta_1\beta_2 x_{i1} x_{i2} + \beta_0^2 + \beta_1^2 x_{i1}^2 + \beta_2^2 x_{i2}^2), \\ &= \sum_i y_i^2 - 2\beta_0 \sum_i y_i - 2\beta_1 \sum_i y_i x_{i1} - 2\beta_2 \sum_i y_i x_{i2} + 2\beta_1\beta_0 \sum_i x_{i1} + \dots \\ &\qquad \cdots + 2\beta_0\beta_2 \sum_i x_{i2} + 2\beta_1\beta_2 \sum_i x_{i1} x_{i2} + n\beta_0^2 + \beta_1^2 \sum_i x_{i1}^2 + \beta_2^2 \sum_i x_{i2}^2. \end{aligned} \tag{62}$$

Cette expression est une forme quadratique à trois variables (β_0, β_1 et β_2) et les coordonnées minimisant la valeur sont données par :

$$\begin{aligned} \beta_0^* &= \frac{2\sum_i y_i - \beta_1 \sum_i x_{i1} - 2\beta_2 \sum_i x_{i2}}{2n}, \\ &= \bar{y} - \beta_1^* \bar{x}_1 - \beta_2^* \bar{x}_2, \end{aligned} \tag{63}$$

$$\begin{aligned} \beta_1^* &= \frac{2\sum_i y_i x_{i1} - 2\beta_0^* \sum_i x_{i1} - 2\beta_2^* \sum_i x_{i1} x_{i2}}{2\sum_i x_{i1}^2}, \\ &= \frac{\sum_i y_i x_{i1} - \beta_0^* \sum_i x_{i1} - \beta_2^* \sum_i x_{i1} x_{i2}}{\sum_i x_{i1}^2}, \end{aligned} \tag{64}$$

$$\begin{aligned} \beta_2^* &= \frac{2\sum_i y_i x_{i2} - 2\beta_0^* \sum_i x_{i2} - 2\beta_1^* \sum_i x_{i1} x_{i2}}{2\sum_i x_{i2}^2}, \\ &= \frac{\sum_i y_i x_{i2} - \beta_0^* \sum_i x_{i2} - \beta_1^* \sum_i x_{i1} x_{i2}}{\sum_i x_{i2}^2}. \end{aligned} \tag{65}$$

L'idée est alors la même que pour le développement à une seule variable, soit de remplacer l'expression trouvée pour β_0, de regrouper les termes et d'ajouter des termes qui sont égaux à zéro pour transformer l'expression identifiée en des termes de variance et de covariance entre les variables. En commençant par

l'équation 64, on obtient :

$$\begin{aligned}
\sum_i x_{i1}^2 \beta_1^* &= \sum_i y_i x_{i1} - \beta_0^* \sum_i x_{i1} - \beta_2^* \sum_i x_{i1} x_{i2}, \\
&= \sum_i y_i x_{i1} - (\bar{y} - \beta_1^* \bar{x}_1 - \beta_2^* \bar{x}_2) \sum_i x_{i1} - \beta_2^* \sum_i x_{i1} x_{i2}, \\
&\qquad \text{(substitution de } \beta_0^*\text{)} \\
\Rightarrow \left(\sum_i x_{i1}(x_{i1} - \bar{x}_1) \right) \beta_1^* &= \sum_i (y_i - \bar{y}) x_{i1} - \beta_2^* \sum_i x_{i1}(x_{i2} - \bar{x}_2), \\
&\qquad \text{(regroupement des termes)} \\
\Rightarrow \left(\sum_i x_{i1}(x_{i1} - \bar{x}_1) - \underbrace{\bar{x}_1 \sum_i (x_{i1} - \bar{x}_1)}_{=0} \right) \beta_1^* &= \sum_i (y_i - \bar{y}) x_{i1} - \underbrace{\bar{x}_1 \sum_i (y_i - \bar{y})}_{=0} - \beta_2^* \sum_i x_{i1}(x_{i2} - \bar{x}_2) + \dots \\
&\qquad \cdots + \underbrace{\beta_2^* \sum_i \bar{x}_1 (x_{i2} - \bar{x}_2)}_{=0}, \\
&\qquad \text{(ajout d'expressions égales à 0)} \\
\Rightarrow \left(\sum_i (x_{i1} - \bar{x}_1)^2 \right) \beta_1^* &= \sum_i (y_i - \bar{y})(x_{i1} - \bar{x}_1) - \beta_2^* \sum_i (x_{i1} - \bar{x}_1)(x_{i2} - \bar{x}_2), \\
&\qquad \text{(réorganisation des termes)} \\
\Rightarrow \underbrace{\frac{1}{n-1} \sum_i (x_{i1} - \bar{x}_1)^2}_{\text{var}(x_1)} \beta_1^* &= \underbrace{\frac{1}{n-1} \sum_i (y_i - \bar{y})(x_{i1} - \bar{x}_1)}_{\text{cov}(x_1, y)} - \beta_2^* \underbrace{\frac{1}{n-1} \sum_i (x_{i1} - \bar{x}_1)(x_{i2} - \bar{x}_2)}_{\text{cov}(x_1, x_2)}, \\
\Rightarrow \beta_1^* &= \frac{\text{cov}(y, x_1) - \beta_2^* \text{cov}(x_1, x_2)}{\text{var}(x_1)}
\end{aligned} \tag{66}$$

En faisant une démarche similaire avec l'équation 65, on peut obtenir l'ensemble du système exprimé sous forme d'équations compréhensibles :

$$\beta_0^* = \bar{y} - \beta_1^* \bar{x}_1 - \beta_2^* \bar{x}_2, \tag{67}$$

$$\beta_1^* = \frac{\text{cov}(y, x_1) - \beta_2^* \text{cov}(x_1, x_2)}{\text{var}(x_1)}, \tag{68}$$

$$\beta_2^* = \frac{\text{cov}(y, x_2) - \beta_1^* \text{cov}(x_1, x_2)}{\text{var}(x_2)}. \tag{69}$$

Débutons la discussion avec la formule d'estimation du coefficient β_0^*. Sa nouvelle formule est une extension naturelle de celle développée à la section précédente. En plus de contenir la formule originale ($\bar{y} - \beta_1^* \bar{x}_1$), on soustrait maintenant la partie additionnelle qui provient de la nouvelle variable ($-\beta_2^* \bar{x}_2$). Une bonne interprétation de la formule pour le coefficient β_0 correspond à la

valeur résiduelle de $\bar{y}$ une fois qu'on a soustrait le produit des coefficients avec la valeur moyenne de chaque variable.

Pour débuter la discussion à propos de β_1^*, il est utile de comparer l'équation 68 avec l'équation 42, soit la solution à β_1^* quand il n'y a qu'une variable. Mettons les deux équations côte-à-côte :

$$\beta_1^* = \frac{\text{cov}(y, x_1)}{\text{var}(x_1)} \text{ (une variable)}, \quad \beta_1^* = \frac{\text{cov}(y, x_1) - \beta_2^* \text{cov}(x_1, x_2)}{\text{var}(x_1)} \text{ (deux variables)}.$$

Dans la première équation, le coefficient β_1^* n'est déterminé au numérateur que par la covariance entre la variable dépendante et l'unique variable indépendante (x_1). Dans le modèle à deux variables, le coefficient est déterminé par cette même covariance, mais à laquelle on soustrait le produit de la covariance entre les deux variables indépendantes et le coefficient β_2^*.

La formule de l'estimateur du coefficient β_1^* peut donc s'interpréter comme suit. Outre la variance au dénominateur (qui ne change pas d'une formule à l'autres), l'ampleur du coefficient est déterminé par la covariance **qui peut seulement provenir** de la variable indépendante associée à ce coefficient. La covariance qui est partagée par les deux variables indépendantes (le terme $\text{cov}(x_1, x_2)$) est éliminé par la soustraction. Les fluctuations statistiques qui sont associées aux mouvements communs des deux variables indépendantes ne peuvent êtres attribués à une variable ou une autre avec certitude. La formule d'estimation du coefficient β_2 a une interprétation identique : il faut soustraire de la covariance entre y et x_2 les comouvements qui peuvent provenir de l'autre variable indépendante (x_1).

On peut comprendre la formule d'estimation par l'analyse graphique des mouvements qui peuvent êtres expliqués par chacune des variables. J'introduis ce graphique à la Figure 11. Le cercle supérieur représente l'information qu'on cherche à expliquer, soit la variable dépendante (y, dans ce cas). Le cercle inférieur et à gauche représente l'information venant de la première variable indépendante (x_1, dans ce cas) qu'on utilise pour expliquer la variable dépendante. Le cercle inférieur de droite représente quant-à-lui l'information provenant de l'autre variable indépendante (x_2). L'intersection entre le cercle y et le cercle x_1 représente donc la partie de y qui peut-être expliquée par la variable x_1. Cette intersection correspond aux sections colorées en rouge et en bleu. L'intersection entre le cercle y et le cercle x_2 représente quant-à-elle la partie de y qui peut-être expliquée par x_2 (soit les sections en bleu et en gris). La partie en bleu peut expliquer la variable dépendante, mais d'un point de vue statistique, il est impossible de savoir si c'est la variable x_2 ou la variable x_1 qui génère le mouvement. En conséquence, les formules d'estimation d'un coefficient n'en tiennent pas compte et ne considèrent que l'information qui est certaine de provenir de la variable associée.

Biais de variable(s) omise(s)

L'analyse de la formule d'estimation du coefficient β_1^* (équation 68) révèle alors quelque chose qui peut-être surprenant à prime abord. Même si les données

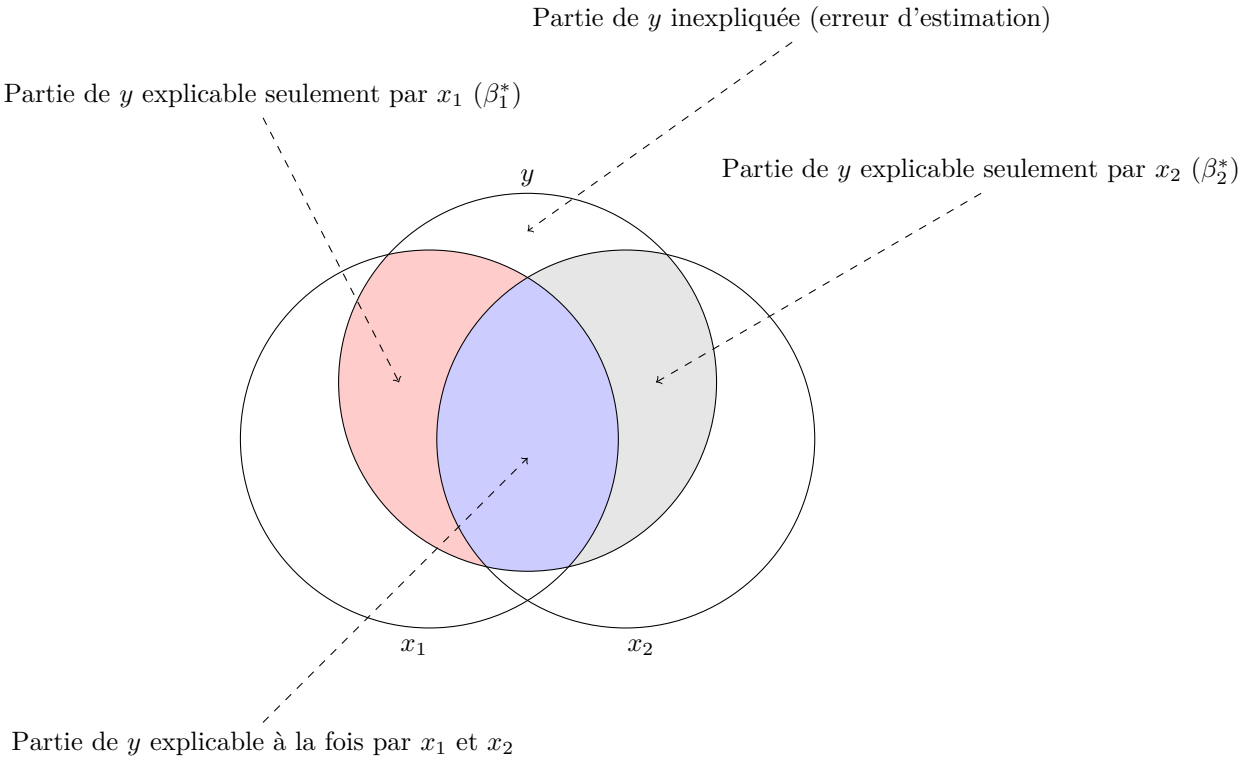

FIGURE 11 – Explication de la formule d'un coefficient β_i^* en termes de comouvements.

échantillonales ne changent pas, l'ajout d'une variable au modèle à estimer changera le coefficient de la variable originale. En effet, l'ajout d'une variable supplémentaire au modèle fait passer l'équation d'estimation de β_1^* de la formule énoncée en 42 à la formule énoncée à l'équation 68. La formule change par l'ajout du terme additionnel $-\beta_2^* \frac{\text{cov}(x_1,x_2)}{\text{var}(x_1)}$. Cet ajout, si la covariance entre les deux variables indépendantes est importante, peut affecter la magnitude... et même faire en sorte que le coefficient change de signe !

En d'autres termes, si on omet de mettre une variable indépendante qui explique la variable dépendante, il y a fort à parier qu'on faussera les coefficients estimés des autres variables du modèle. C'est ce qu'on appelle un **biais de variables omises**. Omettre une variable du modèle peut donc affecter sévèrement l'interprétation qu'on se fait entre les variables d'un modèle. La seule exception est si la covariance entre la variable omise et toutes les autres variables du modèle est nulle. Dans ce cas, omettre la variable n'aura pas d'effet.

0.5.2 Un autre exemple

Avec le développement des formules identifiées aux équations 67 à 68, on peut estimer le modèle de l'équation 60 à partir des données du tableau 5. Pour ce faire, il faut calculer la covariance entre chaque variable, de même que les moyennes de chaque variable. Je rapporte ici ces valeurs sans détailler les calculs :

$$\begin{aligned} \bar{y} &= \frac{280}{6}, & \bar{x}_1 &= \frac{162}{6}, & \bar{x}_2 &= \frac{3}{6}, \\ \text{cov}(y, x_1) &= 24, & \text{cov}(y, x_2) &= 6, & \text{cov}(x_1, x_2) &= 0.2, \\ \text{var}(x_1) &= 7.6, & \text{var}(x_2) &= 0.3. \end{aligned}$$

On peut alors substituer les valeurs associées dans chaque formule des coefficients :

$$\beta_0^* = \frac{280}{6} - \beta_1^* \frac{162}{6} - \beta_2^* \frac{1}{2}, \tag{70}$$

$$\beta_1^* = \frac{1}{7.6} \left(24 - \beta_2^* 0.2\right), \tag{71}$$

$$\beta_2^* = \frac{1}{0.3} \left(6 - \beta_1^* 0.2\right). \tag{72}$$

Ces équations se résolvent alors avec les techniques usuelles d'algèbre pour obtenir la solution à chaque coefficient :

$$\beta_1^* = \frac{600}{30 \cdot 7.6 - 4} \approx 2.6786, \tag{73}$$

$$\beta_2^* = 20 - \frac{400}{30 \cdot 7.6 - 4} \approx 18.214, \tag{74}$$

$$\beta_0^* = \frac{280}{6} - \frac{600}{30 \cdot 7.6 - 4} \frac{162}{6} - \left(20 - \frac{400}{30 \cdot 7.6 - 4}\right) \frac{1}{2} \approx -34.763, \tag{75}$$

En somme, le modèle estimé de l'équation 60 est donné par :

$$y_i = -34.76 + 2.68x_{i1} + 18.21x_{i2}. \tag{76}$$

Interprétation des coefficients estimés

Le coefficient estimé indique de combien la variable dépendante augmentera si la variable indépendante associée au coefficient estimé change de une unité. Par exemple, si la température augmente de un degré Celsius, le modèle estimé ci-dessus nous indique qu'on peut s'attendre à une augmentation estimée de 2.68 personnes à la piscine. Similairement, si on passe d'un jour de semaine ordinaire ($x_{i2} = 0$) à une journée de fin de semaine ($x_{i2} = 1$), on peut s'attendre à une augmentation de 18.21 personnes à la piscine.

En rétrospective, le modèle estimé semble donner raison au sauveteur de la ville : un modèle qui tient compte de la fin de semaine pour expliquer les baignades semble important.

Certains résultats demeurent cependant les mêmes

On peut montrer qu'avec deux variables, la moyenne des termes d'erreur restera nulle. On peut aussi montrer que la covariance entre ces termes d'erreurs et la partie expliquée sera aussi nulle. Finalement, la droite estimée passera toujours par le centre statistique de l'échantillon. En fait, peut importe le nombre de variables indépendantes qui seront utilisées pour fin d'estimation, ces propriétés seront toujours vraies.

0.5.3 Généralisation des formules d'estimation à $k+1$ variables

La courte discussion sur le biais de variable omises peut soulever une inquiétude : comment estimer un modèle de plus de deux variables ? Après-tout, il est fort possible que le prix d'accès à la piscine, la proximité du fleuve ou d'autres variables puissent influencer le nombre de personnes qui décident de se baigner dans une piscine. Comment estimer un modèle qui inclut ces variables additionnelles ? De manière générale, on peut s'intéresser à estimer l'effet de $k-1$ variables indépendantes $x_{i1}, x_{i2}, \ldots x_{ik}$ sur une variable dépendante : y_i via un modèle linéaire :

$$y_i = \beta_0 + \sum_{j=1}^{k} \beta_j x_{ij} + e_i. \tag{77}$$

L'approche pour dériver les formules estimant chaque coefficient β_j^* est la même que dans les sections précédentes :

1. La somme du carré des erreurs de prédiction (SSR) génère une forme quadratique à $k+1$ variables $(\beta_0, \beta_1, \ldots, \beta_k)$ et minimiser les erreurs de prédiction revient à trouver le sommet de cette forme quadratique.

2. Les coordonnées de cette forme quadratique permettent d'identifier des formules pour les « betas » estimés. Des manipulations algébriques identiques à celles des sections précédentes permettent d'exprimer les formules d'estimation en fonction de moyennes, variances et covariances.
3. La résolution de ce système à $k+1$ équations et $k+1$ inconnues (les k coefficients et la constante) se fait par les techniques usuelles d'algèbre.

En fait, on peut montrer avec -non sans patience !- que les formules d'estimation des coefficients seront de la forme suivante :

$$\beta_0^* = \bar{y} - \sum_{j=1}^{k} \beta_j^* \bar{x}_j, \tag{78}$$

$$\beta_l^* = \frac{\text{cov}(y, x_l) - \sum_{j \neq l} \beta_j^* \text{cov}(x_l, x_j)}{\text{var}(x_l)} \; l = 1 \ldots k. \tag{79}$$

La dernière formule, pour le l-ème coefficient, contient un sommation sur l'index « j qui n'est pas égal à l » $\left(\sum_{j \neq l}\right)$. Cette sommation signifie qu'il faut faire la somme sur tous les autres indices (1, 2, ..., k) sauf celle de l'indice l. Bref, il faut enlever la covariance entre x_l et toutes les *autres* variables indépendantes. Comme la dernière phrase le suggère, l'interprétation de la formule est la même que pour le cas à deux variables. La seule différence est qu'il faut faire l'analyse graphique de la Figure 11 pour chacune des $k-1$ autres variables.

0.6 Une première mesure de force estimative d'un modèle

Cette section présente un premier critère de comparaison des modèles d'estimation statistique. Ce critère de comparaison ne reflète en rien les forces ou faiblesses des propriétés statistiques du modèle estimé, mais évalue *seulement* sa performance expliquer l'échantillon. Ce critère se nomme le **R-carré**.[3]

Pour comprendre ce que veut dire le R-carré, commençons par décomposer chaque observation y_i par la somme de ce qui peut-être expliqué par le modèle ($y_i^* \stackrel{\text{def}}{=} \beta_0^* + \sum_{j=1}^{k-1} \beta_j^* x_{ij}$) et ce qui ne peut être expliqué par le modèle

$$y_i = y_i^* + e_i * .$$

Jusqu'ici, rien de sorcier. On peut transformer cette identité en soustrayant la moyenne de chaque côté pour obtenir :

$$(y_i - \bar{y}) = (y_i^* - \bar{y}) + (e_i^*). \tag{80}$$

Souvenez-vous que parce qu'un échantillon ne fait aucune erreur à la moyenne de l'échantillon, la moyenne de y_i^* est aussi $\bar{y}$. Comme nous savons également

3. Selon les textes ou logiciels, on verra aussi la notation R^2 ou « R-deux ». Toute ressemblance avec Star Wars est purement fortuite.

que la moyenne des résidus est nulle ($\bar{e} = 0$) par construction, nous avons que chaque terme entre parenthèses, dans l'expression ci-dessus, correspond à sa déviation par rapport à sa moyenne (car $e_i - 0 = e_i$). Prenons cette identité et élevons là à la puissance deux. On obtient alors :

$$(y_i - \bar{y})^2 = (y_i^* - \bar{y})^2 + 2(y_i^* - \bar{y})e_i + (e_i*)^2. \tag{81}$$

Maintenant, effectuons la sommation sur toutes les observations :

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(y_i^* - \bar{y})^2 + 2\underbrace{\sum_{i=1}^{n}(y_i^* - \bar{y})e_i}_{=0} + \underbrace{\sum_{i=1}^{n}(e_i*)^2}_{=SSR}. \tag{82}$$

Dans les pages précédentes, nous avons montré que la covariance entre la valeur prédite par l'estimation et les résidus était égale à zéro (voir les sections 0.4.1 et 0.5.2). Comme le terme du centre n'est rien d'autre que cette covariance multipliée par $n-1$, on peut aussi conclure qu'il est égal à zéro. Il s'en suit que l'identité développée est la suivante :

$$\underbrace{\sum_{i=1}^{n}(y_i - \bar{y})^2}_{\stackrel{\text{def}}{=} TSS} = \underbrace{\sum_{i=1}^{n}(y_i^* - \bar{y})^2}_{\stackrel{\text{def}}{=} ESS} + \underbrace{\sum_{i=1}^{n}(e_i*)^2}_{=SSR}, \tag{83}$$

$$TSS = ESS + SSR, \tag{84}$$

où les termes « TSS » et « ESS » sont respectivement des abréviations tirées de l'anglais pour *total sum of squares* (somme totale des carrés) et *explained sum of squares* (somme au carrée de la partie expliquée). L'abréviation SSR désigne toujours *sum of square residuals* (somme du carré des résidus).

Cette dernière expression n'est rien d'autre qu'une formulation statistique de l'identité géométrique connue comme le « Théorème de Pythagore ». La somme des carrés expliqués représente l'hypothénuse d'un triangle rectangle, la somme des carrés expliqués correspond à sa base et la somme du carré des résidus correspond à sa hauteur. J'ai représenté cette relation à la Figure 12.

Le R-carré n'est rien d'autre que le ratio du ESS au TSS, soit :

$$R^2 \stackrel{\text{def}}{=} \frac{ESS}{TSS} = 1 - \frac{SSR}{TSS}. \tag{85}$$

Le R-carré mesure le ratio (du carré) de la partie expliquée à l'ensemble (du carré) de la variable dépendante. C'est un nombre compris entre zéro et un. Un R-carré proche de un (1) suggère donc un modèle statistique qui explique très bien la variable dépendante alors qu'un R-carré proche de zéro suggère un modèle d'analyse qui explique mal la variable dépendante. J'illustre cette idée à la Figure 13, où je montre la représentation géométrique de deux modèles (indicés 1 et 2) d'analyse expliquant la même variable dépendante (y). Dans cette figure, j'ai pris bien soin que le TSS (l'hypoténuse du triangle) soit de

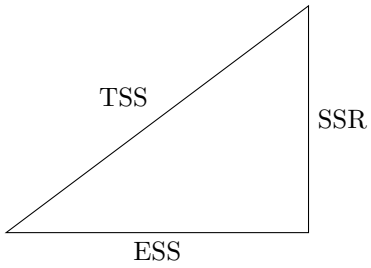

FIGURE 12 – Lien entre le R-carré et le théorème de Pythagore.

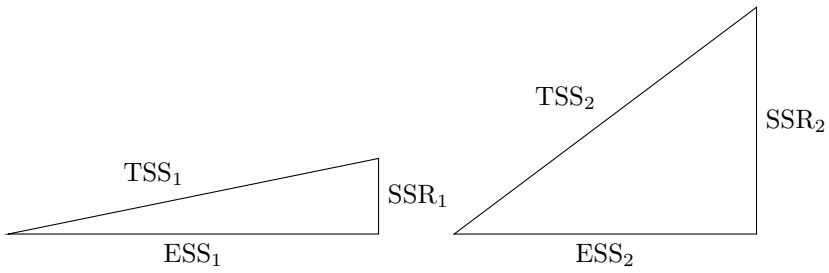

FIGURE 13 – Le modèle 1 explique mieux l'échantillon que le modèle 2.

même longueur pour les deux modèles, reflétant que la variable dépendante ne change pas. Si notre **seul** critère d'analyse de la performance d'un modèle est sa capacité à prédire les données de l'échantillon, il devrait être intuitif pour le lecteur que le premier modèle est beaucoup plus attrayant que le second. En effet, la somme du carré des résidus (hauteur du triangle) est beaucoup plus faible que pour le second modèle. En d'autres termes, la partie prédictive du premier modèle (la base du triangle) arrive beaucoup mieux à prédire la variable dépendante que le second modèle.

Comme le R-carré représente le ratio du ESS au TSS, c'est exactement l'information qu'il traduira. Plus le modèle estimé arrive à bien prédire les données dans l'échantillon, plus il sera proche de un. Le R-carré permet donc d'évaluer rapidement à quel point un modèle arrive bien à prédire la variable dépendante tel qu'échantillonné.

Ce développement peut soulèver trois questions. D'abord, pourquoi le R-carré élimine la moyenne des données observées dans ses calculs ? S'il cherche à expliquer la capacité d'un modèle à expliquer la variable dépendante, pourquoi ne pas prendre les données intégrales, sans soustraire la moyenne ? Ensuite, en supposant qu'un R-carré élevé soit une propriété désirable d'un modèle d'analyse, comment augmenter sa valeur ? Ensuite, une question préalable à la question précédente, ne faut-il pas conclure qu'un modèle avec un R-carré élevé est un *bon modèle* pour faire des prévisions ? Je développe des réponses à ces questions ci-dessous.

L'idée de soustraire la moyenne pour fin d'analyse de performance est très simple : n'importe quel modèle peut au minimum arriver à obtenir la moyenne des observations comme valeur prédites des observations. En effet, il suffit d'utiliser le modèle sans aucune variable $y_i = \beta_0 + e_i$. Nous avons vu, dans les exercices, que le coefficient estimé de ce modèle était alors $\beta_0^* = \bar{y}$. En somme, même le modèle le plus simple arrive à produire cette performance. En conséquence, l'analyse du R-carré se fait uniquement sur le pouvoir prédictif de variables **en plus** de la constante β_0. En soustrayant la moyenne des observations, on ne comptabilise pas indûment l'impact que pourrait avoir la constante sur la capacité prédictive du modèle. On mesure donc la performance du modèle par rapport à un modèle par défaut, ne contenant que la constante.

En ce qui concerne la deuxième question, on peut rendre le R-carré d'un modèle arbitrairement proche de un (1) en ajoutant mécaniquement de nouvelles variables. Intuitivement, il suffit de penser qu'en ajoutant une variable supplémentaire, il est toujours possible de garder le même modèle que précédemment. En conséquence, dès que le coefficient associé à cette nouvelle variable n'est pas égal à zéro, c'est que son ajout permet de réduire un tant soit peu la somme du carré des résidus (le SSR). Bref, ajouter une variable dans une régression ne peut jamais nuire à la valeur du R-carré.

En fait, on peut montrer que pour un échantillon de taille n, il est possible de trouver $n-1$ variables telles que le R-carré sera exactement égal à un (1) et ce, *indépendamment* de la relation statistique entre ces variables. Pour s'en convaincre, rappelez-vous qu'un échantillon de taille n permet d'estimer exactement n coefficients, soit $n-1$ coefficients devant des variables et une constante. Avec un total de n variables, le modèle de prévision devient un système à n équations et n inconnues. Sous certaines conditions mathématiques, un sytème de la sorte admet au plus une seule solution, ce qui revient à dire que le modèle arrive à expliquer parfaitement les données expliquées. [4] Je vais même plus loin : à partir *d'une seule* variable x_1, il est possible, en créant $n-2$ variables basées transformées à partir de cette variable, d'obtenir un R-carré égal à un (1) !

Ces deux paragraphes suggèrent une stratégie pour augmenter le R-carré autant qu'on le souhaite. Si on veut parfaitement expliquer la variable dépendante, il suffit d'ajouter un grand nombre de variables, ou de générer des transformations sur une même variable, et nous obtiendrons un R-carré parfait. Le modèle développé prédirait alors parfaitement la variable dépendante dans l'échantillon.

Cette discussion devrait donner la puce à l'oreille quant à la réponse à la deuxième question. Un bon modèle a-t-il nécessairement un R-carré élevé ? La réponse courte est non. Rappelons-nous que le seule vertu du R-carré est de mesurer la performance du modèle à réduire le SSR *à l'intérieur de l'échantillon.* Or, le but d'un modèle statistique est de prédire la réalité. L'échantillon n'est qu'un outil. Si on effectuait un nouvel échantillon de n' observations, à quel point le modèle original arriverait à bien prédire les nouvelles données obtenues ? Après tout, le but d'un modèle statistique n'est pas de prédire ce qui se passe dans un échantillon déjà connu, mais bien de prédire ce qui se passe dans la

4. La condition nécessaire est l'absence de multi-colinéarité entre les variables.

réalité !

Si un modèle avec un R-carré élevé prédira extrêmement bien la variable dépendante d'un même échantillon, il sera généralement de piètre performance pour faire des prédictions *en dehors* de cet échantillon (c'est-à-dire, dans la réalité). Un modèle avec un R-carré élevé est sur-spécialisé pour l'échantillon. Comme cet échantillon est pris au hasard, la réalisation des variables observées dépend en partie de phénomènes complètement aléatoires. Pour fin de prévision en situation pratique, c'est-à-dire en dehors de l'échantillon, on veut justement que le modèle ne **tienne pas** compte de ces phénomènes aléatoires. On veut qu'il préserve seulement les phénomènes qui se reproduiront dans les échantillons suivants. Lorsque qu'un modèle est surspécialisé pour un échantillon, on dit que le modèle est **suridentifié**.

Pour aider à comprendre, j'emploie une analogie. Supposons, pour fin de discussion, qu'on souhaite développer une bonne technique pour qu'un ordinateur puisse dessiner un arbre. Une approche consiste à prendre une photo pour ensuite entraîner l'ordinateur à dessiner cet arbre à partir de la photo. Il faut comprendre que la photo comprends des particularités qui lui sont propres, mais n'ont pas à voir avec l'arbre. Des vents présents peuvent avoir fait pencher l'arbre d'un côté, la lumière peut être particulière, et ainsi de suite. Il devrait alors être clair que l'ordinateur, qui s'entraîne *exclusivement* à dessiner des arbres à partir de cette photo, sera très bon pour reproduire... la photo. Du moment qu'on lui demande de tracer un autre arbre, il reproduira nécessairement les biais associés avec la photo : l'arbre est penché, la couleur est trop claire et ainsi de suite.

Dans cette analogie, l'arbre représente la réalité, la photo est l'échantillon statistique de cette réalité, l'ordinateur apprenant à dessiner est un analyste choisissant un modèle de régression et les particularité de la photos représentent les phénomènes hasardeux qui sont capturés par l'échantillon, mais qui n'ont rien à voir avec le dessin d'un arbre. L'analogie cherche à illustrer que surspécialiser un modèle pousse se modèle à capturer des phénomènes hasardeux qui n'ont rien à voir avec *l'essence* du phénomène statistique étudié.

Pour construire un bon modèle statistique, il faut aussi tenir compte de la capacité du modèle à faire des prévisions fiables, en dehors de l'échantillon d'estimation. Cette capacité prédictive dans la réalité est généralement contradictoire avec la mesure du R-carré. Pour développer une meilleure compréhension de ce qu'est un « bon » modèle, il nous faut développer l'artillerie statistique sur les notion de hasard et de probabilités, ce qui sera développés dans les prochains chapitres. Nous verrons que les critères utilisés pour développer des bons modèles sont en fait très loin du R-carré.

Il faut cependant rendre à César ce qui lui appartient. Si le R-carré n'est pas une bonne mesure de la qualité d'un modèle statistique, il reste cependant utile pour améliorer la performance dans certaines circonstances. Ces circonstances seront détaillées dans des chapitres ultérieurs, où l'on cherchera justement à évaluer la performance relative de modèles similaires.

0.7 Résumé

Dans ce chapitre, nous avons couvert l'essentiel de la théorie mathématique entourant l'analyse de régression. Nous avons développé une méthode qui permet de trouver les « coefficients mystères » $\beta_0, \beta_1, \ldots, \beta_k$ de la droite de régression :

$$y_i = \beta_0 + \sum_{i=1}^{k} \beta_j x_{ij} + e_i. \tag{86}$$

Nous avons vu que si nous cherchions à minimiser le carré des erreurs de prédiction ($\sum_i e_i^2$), alors les coefficients estimés étaient la solution au système d'équations linaire à $k+1$ variables et $k+1$ équations suivant :

$$\beta_0^* = \bar{y} - \sum_{j=1}^{k} \beta_j^* \bar{x}_j, \tag{87}$$

$$\beta_l^* = \frac{\text{cov}(y, x_l) - \sum_{j \neq l} \beta_j^* \text{cov}(x_l, x_j)}{\text{var}(x_l)} \; l = 1 \ldots k. \tag{88}$$

L'interprétation intuitive de ces équations est que les coefficients associé à une variable dépendante est une fonction de la covariance entre cette même variable dépendante et la variable indépendante, une fois qu'on a éliminé le somme des effets des autres variables dépendantes.

Cette analyse nous aura conduit à parler du concept de biais de variables omises, c'est-à-dire le changement important de la valeur d'un ou plusieurs coefficients de par l'absence d'une variable dans le modèle estimé. Si une telle variable venait à manquer, les coefficients pourraient donner de fausses indications sur la relation entre la variable dépendante et la variable indépendante.

Nous aurons aussi parlé d'un indicateur de « performance » d'un modèle estimé, soit le R-carré. Nous avons parlé du fait que cet indicateur, compris entre zéro et un, indique le pourcentage de la variable dépendante qui est expliqué par les variables indépendantes. En ce sens, un R-carré élevé est désirable. Nous avons cependant discuté du fait qu'un R-carré élevé pouvait être un sérieux problème pour fin de prévision parce que le modèle serait surspécialisé pour l'échantillon que nous avons sous-la main. Un R-carré n'est donc pas une mesure de fiabilité de *prévision*, mais bien une indication que le modèle épouse bien un échantillon.

Si on analyse ces concepts du point de vue de l'analyste, ils sont certainement contradictoires. D'un côté, la notion de biais de variable omise suggère que l'analyste devrait ajouter autant de variables que possible. Ce faisant, il éviterait des phénomènes importants qui affecterait la magnitude (et le signe) des coefficients. D'un autre côté, ajouter trop de variables mène à la suridentification, ce qui rend un modèle surspécialisé par rapport à ses objectifs de prévision. Que faire ?

On peut résoudre ces deux contradictions. Le but des prochains chapitres permettra justement de déterminer si l'ajout d'une ou plusieurs variables affecte

sérieusement les prévisions d'un modèle. Nous consacrerons également du temps à comprendre la notion de *stratégie d'identification*, un concept qui permet justement d'éliminer les potentiels biais de variables qui sont omises.