Fondements statistiques de l'apprentissage automatique

Laurent Risser
CNRS - Institut de Mathématiques de Toulouse (IMT UMR5219)
3IA Artificial and Natural Intelligence Toulouse Institute (ANITI)

ISAE-SUPAERO - 2021/22

Table des matières

Chapitre 1

Introduction

1.1 Modèle linéaire en Sciences de la Décision ?

Motivation

Afin d'étudier quantitativement un phénomène à $p \geq 1$ variables d'entrée $X = \left(X^{(1)}, X^{(2)}, \ldots, X^{(p)}\right)$ et une variable de sortie Y, il est bien pratique de construire un modèle g qui explique par une relation mathématique les valeurs observées de Y en fonction des variables d'entrée :

$$Y = g\left(X^{(1)}, X^{(2)}, \ldots, X^{(p)}\right) .$$

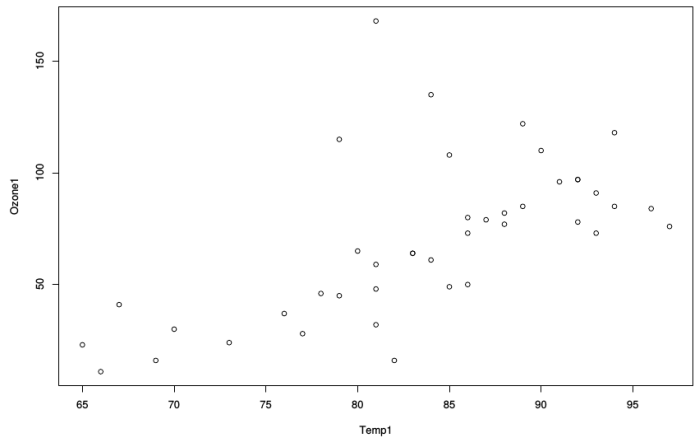
Ce modèle essaie de refléter le plus fidèlement possible la réalité à partir de n observations du phénonène. Il permet ainsi de mieux comprendre le phénomène étudié, mais potentiellement aussi de prédire les sorties inconnues Y en lien avec de nouvelles données d'entrée X.

On distingue deux types de modèles :

1. *Modèles déterministes* : C'est une équation ou un ensemble d'équations qui émanent souvent de lois physiques, chimiques, économiques, ..., et représentent le comportement attendu du phénomène.
2. *Modèles statistiques* : Souvent, il est difficile de développer un modèle théorique car le phénomène étudié est trop complexe. On a alors recours à un modèle statistique basé non pas sur une théorie, mais sur des données observées.

Exemple

On étudie la pollution de l'air à New-York. On a mesuré pendant 111 jours la concentration en ozone, noté 0_i (en ppm), et la température de l'air, notée T_i (en degrés Farenheit). Le tableau ci-dessous représente une partie des observations (celles pour lesquelles la vitesse du vent et le rayonnement solaire sont dans une certaine plage).



On constate que la concentration en ozone croit avec la température. La relation est approximativement linéaire dans la zone représentée ici. En considérant les T_i variables (ou observations) d'entrées et les O_i comme variables (ou observations) de sorties, on introduit alors le modèle :

$$O_i = a + bT_i + \varepsilon_i \,, \tag{1.1}$$

pour chaque observation $i = 1, \cdots, n$, où ε_i représente un bruit entre les observations de sorties réelles et celles prédites par le modèles. Ce modèle est appelé modèle de régression linéaire simple et sera étudié dans ce cours.

Questions posées dans ce cours

La résolution et l'étude du problème introduit ci-dessus sont discutés au début de ce cours (Chapitre 2). Beaucoup d'autres questions permettent de bien comprendre les bases de l'apprentissage statistique, qui est une composante importante de l'Intelligence Artificielle :

- Peut-on s'assurer qu'il y a une relation entre les entrées et les sorties ?
- Quel est le niveau d'incertitude sur cette relation ?
- Peut-on détecter des valeurs abérantes ?
- Que faire si la dimension des entrées (p) est plus grande que le nombre d'observations (n) ?
- Que faire si le niveau de bruit n'est pas le même pour differents groupes de variables ou si différents groupes de variables ont un *bruit* de moyenne non nulle.
- ...

Ces questions seront abordées dand le cadre de ce cour.

1.2 Rappels en Probabilités/Statistique

1.2.1 Notions de variable aléatoire et de densité de probabilité

Variable aléatoire Une *variable aléatoire* (v.a.) X est une application définie sur l'ensemble des résultats possibles d'une expérience aléatoire. Dans le cadre

de ce cours ses résultats possibles seront toujours dans $\mathbb{R}$ ou un sous-ensemble de $\mathbb{R}$. On distinguera en particulier le *cas continu*, par exemple si X représente l'incertitude sur une estimation de la température et le *cas discret*, par exemple $X \in \{0, 1\}$ pour modéliser le résultat lorsque l'on joue à pile ou face.

Loi de probabilité La *loi de probabilité* d'une v.a. décrit la probabilité d'obtenir les différents résultats de cette variable.

Loi de probabilité discrète Par exemple si l'on joue à pile ou face avec une pièce parfaitement équilibrée, on a $\mathbb{P}(X = 0) = 1 - p = 0.5$ et $\mathbb{P}(X = 1) = p = 0.5$. On remarquera que la somme des probabilités de tous les résultats possibles dans le cas discret est toujours 1.

Loi de probabilité continue Dans le cas continu, écrire $\mathbb{P}(X = x)$ n'a aucun sens puisque la probabilité d'une valeur exacte est infinitésimale. On pourra par contre utiliser la *fonction de répartition* $F_X(x) = \mathbb{P}(X \leq x)$ pour représenter comment se répartissent les probabilités des différents résultats de X. Il sera alors possible de quantifier les chances que X soit sur une certaine gamme de valeurs $\mathbb{P}(x_1 < X \leq x_2) = F_X(x_2) - F_X(x_1)$. Naturellement, on aura toujours $F_X(-\infty) = 0$ et $F_X(+\infty) = 1$. De manière purement équivalente à la fonction de répartition $p_X(x)$, la *densité de probabilité* pourra de même représenter la loi de probabilité d'une v.a. X suivant :

$$p_X(x) = \frac{\partial F_X}{\partial x}(x)$$

En utilisant les densités de probabilités, les chances que X tombe sur une gamme de valeurs $[x_1, x_2]$ sera alors

$$\mathbb{P}(x_1 < X \leq x_2) = \int_{x_1}^{x_2} p_X(x)dx\,.$$

.

1.2.2 Théorème central limite

Afin de montrer l'importance de la loi Normale en probabilités/statistique, ainsi que de manipuler les concepts énoncés ci-dessus, il est intéressant de présenter maintenant le Théorème Central Limite (TCL).

Supposons que n variables aléatoires $X_1, X_2, \ldots, X_n$ indépendantes mais suivant une même loi de probabilité soient tirés. L'espérance (ou moyenne) m et l'écart type s de leur loi est connue. Le nombre d'observations n est aussi supposé grand (typiquement $n > 30$). Alors, la somme des X_i peut être approchée par une loi normal de moyenne nm et d'écart type $s\sqrt{n}$, *i.e.* :

$$\sum_{i=1}^{n} X_i \sim \mathcal{N}(nm, s^2 n)\,,$$

où la *densité de probabilité* de la loi normale $\mathcal{N}(\mu, \sigma^2)$ est (voir aussi appendice A) :

$$f_{\theta=\{\mu,\sigma\}}(X_i) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

On peut de même montrer que la loi de $\sum_{i=1}^{n} X_i$ tend de même vers $\mathcal{N}(nm, s^2 n)$ lorsque n tend vers l'infini. Nous ne le montrerons pas ici, mais il est aisé de trouver la preuve de ce théorème.

Afin de nous familiariser avec les notions énoncées ci-dessus, nous proposons de vérifier empiriquement le TCL dans le cas d'une pièce tirée à pile ou face. Le protocole expérimental sera le suivant :

- Chaque étudiant de la classe tire $n = 10$ fois une pièce à pile ou face avec et compte le nombre de fois que la pièce est tombée sur pile. Pile correspond alors à $X_i = 1$ et face à $X_i = 0$.
- On suppose que $\mathbb{P}(X = 1) = 0.5$ et $\mathbb{P}(X = 0) = 0.5$, ce qui est sans doute très proche de la réalité. Ainsi l'espérance (moyenne) de X est $m = 0.5$ et son écart type est $s = 0.5$.
- On va dessiner un graphique dans lequel l'abscisse représente le nombre de 'piles' potentiellement obtenus par un étudiant (entre 0 et 10) et l'ordonnée représente le nombre d'étudiant qui ont obtenus ce nombre de 'piles' divisé par le nombre total d'étudiants.
- On constatera que cette courbe approche la densité de la loi normale de moyenne $10m$ et d'écart type $s\sqrt{10}$ (voir appendice A).

Au delà de la connaissance du TCL lui même et de l'illustration des notions de la section 1.2.1, cet exemple nous amène un enseignement qui est (à mes yeux) l'essence de la modélisation statistique. En assemblant plusieurs variables aléatoires, nous avons créé un modèle aléatoire dont on peut étudier les propriétés statistiques telles que la moyenne mais aussi d'une certaine manière la précision/étendue/sensibilité. Ce type de modélisation se distingue alors de la modélisation déterministe qui ne s'intéresse qu'à l'équivalent de la moyenne ici.

1.2.3 Estimation empirique des paramètres d'un modèle

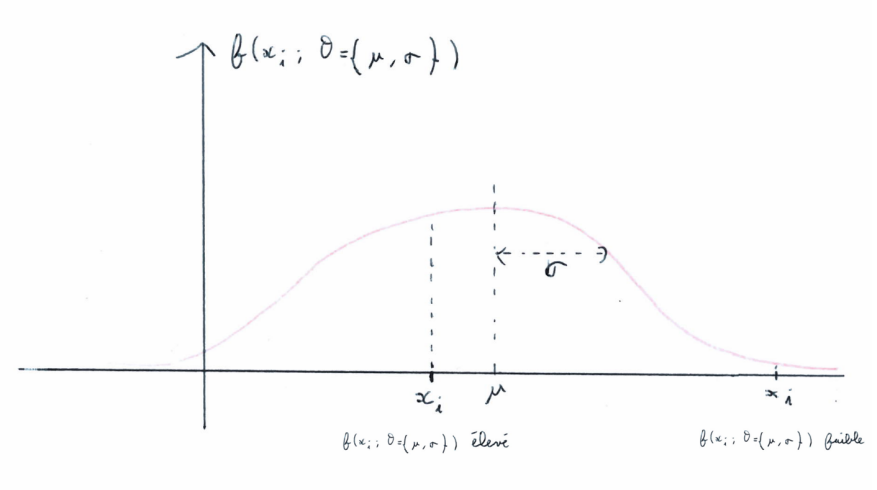
Un des composantes importantes de ce cour est de donner des méthodes pour l'estimation des paramètres de lois à partir d'observations, ou plus spécifiquement de paramètres de modèles contenant des variables aléatoires (c'est à dire avec des sources d'aléa). Cette estimation est classiquement effectuée en suivant le principe du maximum de vraisemblance ou plus simplement une estimation au sens des moindres carrés.

Maximum de vraisemblance

On dénote X une variable aléatoire (v.a.) supposée suivre une loi discrète (e.g. Bernoulli) ou continue (e.g. Normale) de paramètres θ. On note aussi $x_1, \ldots, x_i, \ldots, x_n$ les observations de X.

Pour une observation x_i donnée, on modélise alors la loi de X avec la fonction $f(x_i; \theta)$. Cette fonction vaut $f(x_i; \theta) = \mathbb{P}_\theta(X = x_i)$ si X est une v.a. discrète et $f(x_i; \theta) = f_\theta(x_i)$ si X est continue, où $f_\theta(x_i)$ est la densité de la loi en fonction de ses paramètres θ.

Pour des paramètres θ donnés (ex : moyenne et écart type d'une loi normale), $f(x_i; \theta)$ sera alors d'autant plus élevée que x_i a des chances d'être tirée en fonction des θ.



La vraisemblance des paramètres θ en fonction des observations $x_1, \ldots, x_n$ est alors :

$$L(\theta) = \prod_{i=1}^{n} f(x_i; \theta)$$

Dans l'exemple de pile ou face, supposons que l'on souhaite vérifier empiriquement si une pièce est équilibrée ou non. On modélisera $\mathbb{P}(X = 1) = f(X_i = 1; \theta = \{p\}) = p$ et $\mathbb{P}(X = 0) = f(X_i = 0; \theta = \{p\}) = 1 - p$, puis on réalisera n observations de X en tirant à pile ou face. La vraisemblance sera alors $L(\theta = \{p\}) = \prod_{i=1}^{n}(1_{X_i=1}p + 1_{X_i=0}(1-p))$. Supposons que sur $n = 10$ tirages, on observe 4 'piles' et 6 'faces'. En simplifiant légèrement les notations, la vraisemblance du paramètre p par rapport à notre modèle et nos observations empiriques sera alors $L(p) = p^4(1-p)^6$. Calculons alors la vraisemblance pour plusieurs valeurs de p : $L(0.2) = 0.00042$, $L(0.5) = 0.00098$, $L(0.8) = 0.00002$. De ces trois valeurs, $p = 0.5$ semble le plus vraisemblable.

De manière générale, on calculera le maximum de vraisemblance :

$$\hat{\theta} = \arg\max_{\theta} L(\theta)\,,$$

qui renverra les paramètres les plus vraisemblables en fonction des observations et de la loi choisie.

Dans l'exemple de pile ou face, la meilleur vraisemblance sera obtenue pour $p = 0.4$ avec $L(0.4) = 0.00119$. Si la pièce est bien équilibré, le nombre de 'piles' et de 'faces' obtenus sera de plus en plus proche quand $n \to +\infty$ et $p = 0.5$ aura ainsi la meilleur vraisemblance.

Pour des raisons numériques, il est aussi bien pratique de maximiser la log-

vraisemblance au lieu de la vraisemblance brute :

$$\begin{aligned}\hat{\theta} &= \arg\max_{\theta} \log\left(L(\theta)\right) \\ &= \arg\max_{\theta} \log\left(\prod_{i=1}^{n} f(x_i;\theta)\right) \\ &= \arg\max_{\theta} \sum_{i=1}^{n} \log f(x_i;\theta)\end{aligned}$$

Vu que la fonction *log* est strictement croissante les paramètres optimum $\hat{\theta}$ seront les mêmes avec la log-vraisemblance ou la vraisemblance.

Estimation au sens des moindres carrés

On suppose disposer d'observations $\{y_i\}_{i=\{1,\dots,n\}}$ que l'on souhaite prédire/deviner à partir de observations correspondantes $\{x_i\}_{i=\{1,\dots,n\}}$, où chaque y_i correspond à x_i (voir l'exemple introductif par exemple). Dans ce cours, et très souvent en apprentissage automatique, on va alors optimiser les paramètres θ d'un modèle f_θ pour prédire au mieux les y_i avec $\hat{y_i} = f_\theta(x_i)$.

Faisons l'hypothèse que les erreurs d'approximation du modèle $e_i = y_i - f_\theta(x_i)$ suivent une loi normale centrée, *i.e.* $e_i \sim \mathcal{N}(0,\sigma)$. Ce choix par défaut est commun et semble raisonnable quand f_θ est bien calibré. Nous pouvons alors utiliser le principe de maximum de vraisemblance pour estimer les paramètres θ du modèle f_θ.

$$\begin{aligned}\hat{\theta} &= \arg\max_{\theta} \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{e_i^2}{2\sigma^2}\right) \\ &= \arg\max_{\theta} \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n} e_i^2\right) \\ &= \arg\min_{\theta} \sum_{i=1}^{n} e_i^2 \\ &= \arg\min_{\theta} \sum_{i=1}^{n} (y_i - f_\theta(x_i))^2\end{aligned}$$

Cette technique d'estimation est celle dite au sens des moindres carrés. Nous la retrouvons très couramment en apprentissage automatique et son interprétation est particulièrement intuitive. Elle doit notamment sa popularité au fait qu'il est aisé de calculer son gradient par rapport aux paramètres θ si on sais calculer le gradient de f_θ par rapport à θ :

$$\nabla_\theta e_i^2 = 2(y_i - f_\theta(x_i))\nabla_\theta f_\theta(x_i)$$

Cela ouvre la porte aux techniques d'optimisation par descente de gradient qui sont quasi systématiques en apprentissage automatique.

Pour un public avisé, il faudra se souvenir du fait que la pertinence de l'estimation de paramètres d'un modèle au sens des moindres carrés repose sur une hypothèse de normalité de l'erreur.

Chapitre 2

Regression Linaire

2.1 Régression Linéaire simple

2.1.1 Modèle

On note Y la variable aléatoire réelle à expliquer (ou encore de réponse, dépendante) et X la variable explicative (ou encore déterministe, de contrôle) ou effet fixe ou facteur contrôlé. Le modèle revient à supposer, qu'en moyenne, l'estimation $\mathbb{E}(Y)$, est une fonction affine de X.

$$\mathbb{E}(Y) = f(X) = \beta_0 + \beta_1 X \,.$$

Pour une séquence d'observations aléatoires identiquement distribuées $\{(y_i, x_i)\,, i = 1, \ldots, n\}$, avec $n > 2$ et les x_i non tous égaux, le modèle s'écrit à partir des observations :

$$y_i = \beta_0 + \beta_1 x_i + u_i \,, i = 1, \ldots, n$$

ou bien sous forme matricielle :

$$\begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + \begin{bmatrix} u_1 \\ \vdots \\ u_n \end{bmatrix},$$
$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}$$

où le vecteur u contient les erreurs.

Les hypothèses relatives à ce modèle sont les suivantes :
— la distribution de l'erreur $\mathbf{u}$ est indépendante de X ou bien X est fixe.
— l'erreur est centrée et de variance constante (homoscédasticité) :

$$\forall i = 1, \ldots, n : E(u_i) = 0 \,, Var(u_i) = \sigma_u^2.$$

— β_0 et β_1 sont constants, il n'y a pas de rupture du modèle.
— Hypothèse complémentaire pour les inférences : $u \sim \mathcal{N}(0, \sigma_u^2 \mathbf{I}_p)$. Ce point important est développé dans l'appendice A

Remarque On a fait une hypothèse de linéarité ici mais en pratique cette hypothèse n'est pas toujours valide. Quand ce n'est pas le cas, il existe aussi des méthodes de régression non-paramétriques qui ne sont pas abordées dans le cours mais peuvent être très utiles. Il est aussi possible d'effectuer des transformations élémentaires sur les données, comme par exemple $y_i = \beta_0 + \beta_1 \ln x_i$ ou bien $y_i = \beta_0 + \beta_1 (x_i)^\alpha$.

2.1.2 Estimation

L'estimation des paramètres β_0, β_1, σ_u^2 peut être obtenue en minimisant la somme des carrés des écarts entre observations et modèle (moindres carrés). Pour un jeu de données $\{(y_i, x_i), i = 1, \ldots, n\}$, le critère des moindres carrés s'écrit :

$$\min_{\beta_0, \beta_1} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

Pour minimiser ce critère, on pose :

$$\begin{aligned}
\bar{x} &= \frac{1}{n} \sum_{i=1}^{n} x_i \\
\bar{y} &= \frac{1}{n} \sum_{i=1}^{n} y_i \\
s_x^2 &= \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2 \\
s_y^2 &= \frac{1}{n-1} \sum_{i=1}^{n} (y_i - \bar{y})^2 \\
s_{xy} &= \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) \\
r &= \frac{s_{xy}}{s_x s_y}
\end{aligned}$$

On peut alors montrer que les estimateurs de β_0 et β_1 au sens des moindres carrés sont :

$$\begin{aligned}
b_1 &= \frac{s_{xy}}{s_x^2}, \\
b_0 &= \bar{y} - b_1 \bar{x}.
\end{aligned}$$

On montre que ce sont des estimateurs sans biais et de variance minimum parmi les estimateurs fonctions linéaires des y_i. Cela signifie que pour $i \in \{0, 1\}$ alors $Biais(b_i) = \mathbb{E}(b_i) - \beta_i = 0$ et ainsi que $Var(b_i) = \mathbb{E}(b_i - \mathbb{E}(b_i)) = \mathbb{E}(b_i - \beta_i)$ est minimum. À chaque valeur x_i de X correspond la valeur estimée (ou prédite, ajustée) de Y :

$$\hat{y}_i = b_0 + b_1 x_i$$

les résidus calculés ou estimés sont :

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

La variance σ_u^2 est enfin estimée par la variation résiduelle :

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n} e_i^2 \,.$$

2.1.3 Prédiction

Une fois les paramètres β_0 et β_1 estimés par b_0 et b_1, il est immédiat de prédire la valeur $\hat{y_0}$ qui a le plus de chance d'être associée à une observation x_0 avec :

$$\hat{y_0} = b_0 + b_1 x_0 \,.$$

Il est important de remarquer que le principe d'**estimation** des paramètres d'un modèle à partir de données d'apprentissage (les x_i et y_i) puis de **prédiction** de *scores/labels/variables de sortie* (ici y_0) à partir de nouvelles observations (ici x_0) est au coeur de l'apprentissage automatique.

2.1.4 Inférence

Niveau d'incertitude lié à l'estimation de b_0 et b_1

On rappel qu'une hypothese a été faite sur les résidus $e \sim \mathcal{N}(0, \sigma_u^2 \mathbf{I}_p)$ dans la sous-section 2.1.1 (où e est noté u). Les estimateurs b_0 et b_1 sont alors des variables aléatoires réelles. Ils ne font qu'approcher les valeurs β_0 et β_1 que l'on connaîtrait à coup sur si on disposait d'une infinité d'observations (ou si l'on contrôle le modèle). Ceci est intuitivement évident, si on compare les b_0 et b_1 obtenus sur disons 4 observations pour lesquelles e est faible avec ceux obtenus sur 3 observations avec e faible et une derniere où e est grand, ce qui peut arriver puisque $e \sim \mathcal{N}(0, \sigma_u^2 \mathbf{I}_p)$. Les valeurs de b_0 et b_1 seront différentes alors que le modèle est ses paramètres sont les mêmes.

Sous l'hypothèse de Gaussianité des résidus, on montre que

$$\frac{(n-2)s^2}{\sigma_u^2} \sim \chi^2_{(n-2)}$$

où la loi du χ^2 suit une densité de probabilité donnée appendice A. Alors, les statistiques

$$(b_0 - \beta_0) \Big/ s\left(\frac{1}{n} + \frac{\bar{x}^2}{(n-1)s_x^2}\right)^{1/2}$$

et

$$(b_1 - \beta_1) \Big/ s\left(\frac{1}{(n-1)s_x^2}\right)^{1/2}$$

suivent des lois de Student à $(n-2)$ degrés de liberté. Ceci permet de tester l'hypothèse de nullité d'un de ces paramètres à partir de tests d'hypothèses. On va par exemple tester si le b_1 obtenu est significativement différent de 0, en fonction d'un coefficient α qui représente la probabilité avec laquelle on accepte de se tromper. Typiquement α correspond à 5% de chances de se tromper, ci

qui est raisonablement faible (voir le cours de Statistique pour aller plus loin). Notons, que si b_1 est significativement différent de 0, on peut considérer qu'il existe une relation de dépendance entre les x_i et les y_i.

Intervalles de confiance

Il est de même possible de construire des intervalles de confiance pour les valeurs de b_0 et b_1, toujours en fonction d'un niveau de confiance dépendant de α :

$$b_0 \pm s \left(\frac{1}{n} + \frac{\bar{x}^2}{(n-1)s_x^2} \right)^{1/2} t_{n-2}(\alpha/2)$$
$$b_1 \pm s \left(\frac{1}{(n-1)s_x^2} \right)^{1/2} t_{n-2}(\alpha/2)$$

où $t_\nu(\alpha)$ est la distribution de Student à ν degrés de liberté (voir appendice A). En observant bien ces intervales de confiance ainsi que les distributions de Student, il est interessant de noter que plus on a d'observations n, plus les intervales de confiances sont resserés autour des b_0 et b_1 estimés. Plus on dispose d'information, moins le risque d'erreur est en effet grand par rapport aux valeurs réelles.

Attention : une inférence conjointe sur β_0 et β_1 ne peut être obtenue en considérant séparément les intervalles de confiance. La région de confiance est en effet une ellipse d'équation :

$$n(b_0-\beta_0)^2 + 2(b_0-\beta_0)(b_1-\beta_1)\sum_{i=1}^{n} x_i + (b_1-\beta_1)^2 \sum_{i=1}^{n} x_i^2 = 2s^2 \mathcal{F}_{\alpha;2,(n-2)}$$

où $\mathcal{F}_{\alpha;d_1,d_2}$ et la distribution de Fisher–Snedecor avec les paramètres d_1 et d_2 (voir appendice A).

Niveau d'incertitude lié à l'estimation d'un y_0 à partir d'un x_0

Enfin, connaissant une valeur x_0, on définit deux intervalles de confiance de prédiction à partir de la valeur prédite $\hat{y_0} = b_0 + b_1 x_0$. Le premier encadre E(Y) sachant $X = x_0$; le deuxième, encadre y_0 et est plus grand car il tient compte de la variance totale $\sigma_u^2 + Var(\hat{y_0})$:

$$\widehat{y_0} \pm t_{\alpha/2;(n-2)} s \left(\frac{1}{n} + \frac{(x_0-\bar{x})^2}{(n-1)s_x^2} \right)^{1/2},$$
$$\widehat{y_0} \pm t_{\alpha/2;(n-2)} s \left(1 + \frac{1}{n} + \frac{(x_0-\bar{x})^2}{(n-1)s_x^2} \right)^{1/2}.$$

2.1.5 Qualité d'ajustement

On rappelle que la variance σ_u^2 est estimée par la variation résiduelle :

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n} e_i^2 = \frac{1}{n-1}\sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \frac{1}{n-1}\sum_{i=1}^{n} (y_i - b_0 + b_1 x_i)^2 .$$

et que :

$$s_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \left(\frac{1}{n}\sum_{i=1}^{n}x_i\right)\right)^2$$
$$s_y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i-\bar{y})^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(y_i - \left(\frac{1}{n}\sum_{i=1}^{n}y_i\right)\right)^2$$
$$s_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y}) = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \left(\frac{1}{n}\sum_{i=1}^{n}x_i\right)\right)\left(y_i - \left(\frac{1}{n}\sum_{i=1}^{n}y_i\right)\right)$$

Dans l'optique de mesurer la qualité d'ajustement du modèle, il est d'usage de décomposer les sommes de carrés des écarts à la moyenne sous la forme ci-dessous :

— Sum of Squares Total : $SST = (n-1)s_y^2$
— Sum of Squares Regression : $SSR = (n-1)\frac{s_{xy}^2}{s_x^2}$
— Sum of Squares Errors : $SST = (n-1)s^2$

On appelle alors *coefficient de détermination* la quantité :

$$R^2 = r^2 = \frac{s_{xy}^2}{s_x^2 s_y^2} = 1 - \frac{s^2}{s_y^2} = \frac{SSR}{SST}$$

qui exprime le rapport entre la variance expliquée par le modèle et la variance totale. En pratique, si R^2 vaut par exemple 0.79, cela signifie que 79% de la variablilité de Y a été capturée par le modèle linéaire et que seulement 21% restent à expliquer.

2.1.6 Détection d'outliers

Le critère des moindres carrés est très sensible à des observations atypiques hors "norme" (outliers) c'est-à-dire qui présentent des valeurs trop singulières. L'étude descriptive initiale permet sans doute déjà d'en repérer mais c'est insuffisant. Un diagnostic doit être établi dans le cadre spécifique du modèle recherché afin d'identifier les observations influentes c'est-à-dire celles dont une faible variation du couple (x_i, y_i) induisent une modification importante des caractéristiques du modèle.

Ces observations repérées, il n'y a pas de remède universel : supprimer une valeur aberrante, corriger une erreur de mesure, construire une estimation robuste (en norme L_1), ne rien faire... , cela dépend du contexte et doit être négocié avec le commanditaire de l'étude.

Effet levier

Une première indication est donnée par l'éloignement de x_i par rapport à la moyenne $\bar{x}$. En effet, écrivons les prédicteurs y_i comme combinaisons linéaires des observations :

$$\hat{y}_i = b_0 + b_1 x_i = \sum_{j=1}^{n} h_{ij} y_j$$

avec

$$h_{ij} = \frac{1}{n} + \frac{(x_i - \bar{x})(x_j - \bar{x})}{\sum_{k=1}^{n}(x_k - \bar{x})^2}$$

en notant $\mathbf{H}$ la matrice (hat matrix) des h_{ij} ceci s'exprime encore matriciellement :

$$\hat{\mathbf{y}} = \mathbf{H}\mathbf{y}$$

Les éléments diagonaux h_{ii} de cette matrice mesurent ainsi l'impact ou l'importance du rôle que joue y_i dans l'estimation de $\hat{y_i}$.

Résidus

Différents types de résidus sont définis afin d'affiner leurs propriétés :
— Résidus : $e_i = y_i - \widehat{y_i}$
— Résidus$_i$: $e_{(i)i} = y_i - \widehat{y_{(i)i}} = \frac{e_i}{1-h_{ii}}$
où $\widehat{y_{(i)i}}$ est la prévision de y_i calculée sans la $i$ème observation (x_i, y_i).

Afin de supprimer l'influence de la variance dans les résidus, on remarque d'abord que $Var(e_i) = \sigma_u^2(1 - h_{ii})$. En supposant que $E(e_i) = 0$, les résidus peuvent alors être standardisés de deux manières. Les *résidus standardisés* r_i sont calculés avec :

$$r_i = \frac{e_i}{s\sqrt{1 - h_{ii}}}\,.$$

La standardisation ci-dessus dépend cependant de e_i dans le calcul de s (qui estime $Var(e_i)$). Une estimation non biaisée de cette variance est basée sur

$$s_{(i)}^2 = \left((n-1)s^2 - \frac{e_i^2}{1 - h_{ii}}\right)/(n-3)$$

qui ne tient pas compte de la $i$ème observation. On définit alors les *résidus studentisés* par :

$$t_i = \frac{e_i}{s_{(i)}\sqrt{1 - h_{ii}}}$$

Sous hypothèse de normalité, on montre que ces résidus suivent une loi de Student à $(n-3)$ degrés de liberté.

Il est ainsi possible de construire un test d'hypothèse pour tester la présence d'observations atypique. Plusieurs observations peuvent de même être simultanément considérées en utilisant l'inégalité de Bonferroni. En pratique, les résidus studentisés sont sonvent comparés aux bornes ± 2. Si un résidu studentisé n'est pas dans cet intervale de valeurs, il est considéré comme atypique.

Diagnostics

Un dernier indicateur couramment utilisé est la distance de Cook :

$$D_i = \frac{\sum_{j=1}^{n}(\widehat{y_{(i)j}} - \widehat{y_j})^2}{2s^2} = \frac{h_{ii}}{2(1 - h_{ii})} r_i^2\,, \forall i$$

qui mesure l'influence de chaque observation i sur l'ensemble des prévisions en prenant en compte effet levier et importance des résidus.