La droite des moindres carrés

1 Objectifs

Observer le nuage constitué de n points de coordonnées $(x_i; y_i)$. Si une relation linéaire semble exister (les points ont l'air de se répartir sur une droite), trouver l'équation $y = ax+b$ de la droite qui approche ou « **ajuste** » au mieux ce nuage **au sens des moindres carrés**.

Calculer une incertitude expérimentale sur la pente a de la droite trouvée sous l'hypothèse qu'il existe vraiment une relation linéaire entre x et y.

2 Méthode des moindres carrés

Soit $\boxed{y = ax + b}$ l'équation de la droite cherchée.
Il faut trouver sa pente a et son ordonnée à l'origine b afin de **minimiser la somme des carrés des écarts verticaux** :

$$\sum_{i=1}^{n} (y_i - (ax_i + b))^2 = \sum_{i=1}^{n} (y_i - \widehat{y_i})^2 \qquad \text{en notant : } \boxed{\widehat{y_i} = ax_i + b}$$

$\boxed{e_i = y_i - \widehat{y_i}}$ est l'écart vertical entre le point $(x_i; y_i)$ du nuage et le point correspondant $(x_i; \widehat{y_i})$ sur la droite de régression (d'abscisse aussi égale à x_i).
L'écart vertical e_i est souvent appelé **écart résiduel** ou **résidu**.
e_i est ainsi l'écart entre la valeur observé y_i et sa valeur estimée $\widehat{y_i} = ax_i + b$ connaissant x_i.
La méthode des moindres carrés permet de minimiser $\sum_{i=1}^{n} e_i^2$.

3 Calculs de a et b

$$\begin{aligned}\sum_{i=1}^{n} e_i^2 &= \sum_{i=1}^{n} \left(y_i^2 - 2y_i\,(ax_i + b) + (ax_i + b)^2\right) \\ &= \sum_{i=1}^{n} y_i^2 - 2a\sum_{i=1}^{n} x_i y_i - 2b\sum_{i=1}^{n} y_i + a^2\sum_{i=1}^{n} x_i^2 + 2ab\sum_{i=1}^{n} x_i + \sum_{i=1}^{n} b^2 \\ &= n\cdot\left(\overline{y^2} - 2a\overline{xy} - 2b\overline{y} + a^2\overline{x^2} + 2ab\overline{x} + b^2\right)\end{aligned}$$

Considérons a constant, $\sum_{i=1}^{n} e_i^2$ est l'expression d'une fonction du second degré ayant pour unique variable b, plus précisément :

$$\frac{1}{n}\cdot\sum_{i=1}^{n} e_i^2 = b^2 - 2(\overline{y} - a\overline{x})b + \overline{y^2} - 2a\overline{xy} + a^2\overline{x^2}$$

Son minimum est atteint quand $b = -\dfrac{-2(\overline{y} - a\overline{x})}{2\cdot 1} = \overline{y} - a\overline{x}$ (abscisse du sommet).

Ainsi : $\boxed{b = \overline{y} - a\overline{x}}$. Pour une valeur de a fixée quelconque (la « bonne » valeur n'est pas encore connue) on sait calculer b pour minimiser $\sum_{i=1}^{n} e_i^2$.
Remarquons que la formule trouvée pour b nous indique que le point moyen $(\overline{x};\overline{y})$ doit appartenir à la droite cherchée !

Cherchons a et pour cela considérons b constant, $\sum_{i=1}^{n} e_i^2$ est l'expression d'une fonction du second degré ayant pour unique variable a, plus précisément :

$$\frac{1}{n}\bullet\sum_{i=1}^{n} e_i^2 = \overline{x^2}a^2 + 2(\overline{x}b - \overline{xy})a + b^2 - 2\overline{y}b + \overline{y^2}$$

Son minimum est atteint quand $a = -\dfrac{2(\overline{x}b - \overline{xy})}{2\bullet\overline{x^2}} = \dfrac{-\overline{x}(\overline{y} - a\overline{x}) + \overline{xy}}{\overline{x^2}} = \dfrac{\overline{xy} - \overline{x}\bullet\overline{y}}{\overline{x^2}} + a\dfrac{\overline{x}\bullet\overline{x}}{\overline{x^2}}$

On obtient donc : $a(1 - \dfrac{\overline{x}\bullet\overline{x}}{\overline{x^2}}) = \dfrac{\overline{xy} - \overline{x}\bullet\overline{y}}{\overline{x^2}}$ soit $a = \dfrac{\overline{xy} - \overline{x}\bullet\overline{y}}{\overline{x^2} - \overline{x}\bullet\overline{x}}$

4 Variances et covariances observées

Le dénominateur de a est la variance observée $s_x^2 = \overline{x^2} - \overline{x}\bullet\overline{x} = \dfrac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2$

Le numérateur de a est la covariance observée $s_{xy} = \overline{xy} - \overline{x}\bullet\overline{y} = \dfrac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})\bullet(y_i - \overline{y})$

Donc $\boxed{a = \dfrac{\overline{xy} - \overline{x}\bullet\overline{y}}{\overline{x^2} - \overline{x}\bullet\overline{x}} = \dfrac{s_{xy}}{s_x^2}}$ et l'équation de la droite est $\boxed{y - \overline{y} = \dfrac{s_{xy}}{s_x^2}\bullet(x - \overline{x})}$

5 Moyenne et variance observées des résidus

Connaissant x_i, on peut estimer y_i en calculant $\widehat{y_i} = ax_i + b = \overline{y} + \dfrac{s_{xy}}{s_x^2}\bullet(x_i - \overline{x})$

$\boxed{\text{La moyenne observée des résidus est nulle : } \overline{e} = 0}$. Les n résidus ne sont pas indépendants. En effet :

$$\begin{aligned}\overline{e} &= \frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{y_i}) \\ &= \frac{1}{n}\sum_{i=1}^{n}(y_i - (ax_i + b)) \\ &= \frac{1}{n}\sum_{i=1}^{n} y_i - \frac{a}{n}\sum_{i=1}^{n} x_i - \frac{b}{n}\sum_{i=1}^{n} 1 \\ &= \overline{y} - (a\overline{x} + b) \\ &= \overline{y} - \overline{y} \\ &= 0\end{aligned}$$

Ainsi, la variance observée des résidus est $s_e^2 = \overline{e^2} - \overline{e}^2 = \overline{e^2}$

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n} e_i^2 &= \frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{y_i})^2 \\ &= \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \left(\overline{y} + \frac{s_{xy}}{s_x^2}\bullet(x_i - \overline{x})\right)\right)^2 \\ &= \frac{1}{n}\sum_{i=1}^{n}\left((y_i - \overline{y}) - \frac{s_{xy}}{s_x^2}\bullet(x_i - \overline{x})\right)^2 \\ &= s_y^2 - 2\frac{s_{xy}^2}{s_x^2} + \frac{s_{xy}^2}{s_x^4}\bullet s_x^2 \\ &= s_y^2\bullet\left(1 - \frac{s_{xy}^2}{s_x^2\bullet s_y^2}\right) \\ &= \boxed{s_y^2\bullet\left(1 - r^2\right) = s_e^2}\end{aligned}$$

6 Coefficient de détermination : r^2

L'expression de la variance s_e^2 des résidus e_i permet d'introduire le **coefficient de détermination** r^2 et le **coefficient de corrélation** : $\boxed{r = \frac{s_{xy}}{s_x\bullet s_y}}$

Si $r^2 = 1$, alors on a $y_i = \widehat{y_i}$ pour les n points, qui sont donc parfaitement alignés. La connaissance de x_i permet de déterminer la valeur de y_i.

De façon générale, plus r^2 est proche de 1, plus la variance observée σ_e^2 des écarts verticaux est proche de 0, plus les estimations $\widehat{y_i}$ connaissant x_i sont proches des valeurs observés y_i.

Par ailleurs, on trouve que la pente de la droite des moindres carrés $a = \frac{s_{xy}}{s_x^2} = \frac{r\bullet s_x\bullet s_y}{s_x^2} = \boxed{r\bullet\frac{s_y}{s_x}}$

Ainsi, si r=1, la pente de la droite de régression est le rapport $\frac{s_y}{s_x}$, ce qui est cohérent avec la définition de la pente (rapport des variations en y sur les variations en x).

7 Incertitude expérimentale sur l'estimation a de la pente α

On suppose que $y = \alpha x + \beta$

La pente a de la droite de régression est une estimation de α obtenue à partir de n observations indépendantes $(x_i; y_i)$ (i variant de 1 à n). Remarquons que l'expression de a peut aussi s'écrire :

$$a = \frac{s_{xy}}{s_x^2} = \frac{\sum_{i=1}^n (x_i - \overline{x}) \cdot (y_i - \overline{y})}{\sum_{i=1}^n (x_i - \overline{x})^2} = \boxed{\frac{\sum_{i=1}^n (x_i - \overline{x}) \cdot y_i}{\sum_{i=1}^n (x_i - \overline{x})^2} = a}$$

En effet, $\sum_{i=1}^n (x_i - \overline{x}) \cdot (y_i - \overline{y}) = \sum_{i=1}^n (x_i - \overline{x}) \cdot y_i - \overline{y} \cdot \sum_{i=1}^n (x_i - \overline{x}) = \sum_{i=1}^n (x_i - \overline{x}) \cdot y_i - \overline{y} \cdot 0$

FIXONS les x_i.

Aux n observations $(x_i; y_i)$, il correspond n variables aléatoires $Y_i = \alpha x_i + \beta + E$.

On définit ainsi la variable aléatoire E, d'espérance mathématique (moyenne) $\mathrm{E}[E] = 0$ et de variance $\mathrm{Var}(E) = \sigma^2$, qui modélise l'erreur de mesure sur y, supposée indépendante de la valeur de x_i.

La valeur de x_i étant fixée, on a $\mathrm{E}[Y_i] = \alpha x_i + \beta + \mathrm{E}[E] = \alpha x_i + \beta$ et $\mathrm{Var}(Y_i) = \mathrm{Var}(E) = \sigma^2$.

L'estimation a de la pente α est donc une réalisation de la variable aléatoire $A = \dfrac{\sum_{i=1}^n (x_i - \overline{x}) \cdot Y_i}{\sum_{i=1}^n (x_i - \overline{x})^2}$

$$\begin{aligned}
\mathrm{E}[A] &= \frac{\sum_{i=1}^n (x_i - \overline{x}) \cdot \mathrm{E}[Y_i]}{\sum_{i=1}^n (x_i - \overline{x})^2} = \frac{\sum_{i=1}^n (x_i - \overline{x}) \cdot (\alpha x_i + \beta)}{\sum_{i=1}^n (x_i - \overline{x})^2} \\
&= \frac{\alpha \sum_{i=1}^n x_i^2 - \alpha \overline{x} \sum_{i=1}^n x_i + \beta \sum_{i=1}^n (x_i - \overline{x})}{\sum_{i=1}^n (x_i - \overline{x})^2} = \alpha \cdot \frac{\sum_{i=1}^n x_i^2 - \overline{x} \sum_{i=1}^n x_i}{\sum_{i=1}^n (x_i - \overline{x})^2} \\
&= \alpha \cdot \frac{n\overline{x^2} - n\overline{x}^2}{n \cdot s_x^2} = \alpha
\end{aligned}$$

Donc a est une estimation sans biais de α.

$$\mathrm{Var}(A) = \frac{\sum_{i=1}^n (x_i - \overline{x})^2 \cdot \mathrm{Var}(Y_i)}{\left(\sum_{i=1}^n (x_i - \overline{x})^2\right)^2} = \sigma^2 \cdot \frac{\sum_{i=1}^n (x_i - \overline{x})^2}{\left(\sum_{i=1}^n (x_i - \overline{x})^2\right)^2} = \frac{\sigma^2}{\sum_{i=1}^n (x_i - \overline{x})^2} = \frac{\sigma^2}{n \cdot s_x^2}$$

Si on suppose maintenant que l'erreur de mesure E suit une loi normale de moyenne 0 et d'écart type σ, alors on en déduit immédiatement que les Y_i suivent une loi normale de moyenne $\alpha x_i + \beta$ et d'écart type σ et aussi que A suit une loi normale de moyenne α et d'écart type $\dfrac{\sigma}{\sqrt{n} \cdot s_x}$.

Le problème est qu'on ne connaît pas σ. Là encore, comme pour les incertitudes expérimentales, nous allons construire une loi de Student formée du rapport de deux lois ayant chacune une variance qui dépend de σ. Le quotient de ces variances fera disparaître σ.

Supposons que l'on a effectué n mesures indépendantes pour les y_i, alors on trouve ainsi que $\dfrac{(a - \alpha)\, s_x}{\frac{s_e}{\sqrt{n-2}}} = \dfrac{a - \alpha}{\frac{\sqrt{1-r^2}}{\sqrt{n-2}} \frac{s_y}{s_x}}$ suit une loi de Student à n-2 degrés de liberté.

En notant t_{n-2} le **coefficient de Student** pour un intervalle de 95 %, on en déduit immédiatement l'incertitude correspondante : $\boxed{t_{n-2} \cdot \frac{\sqrt{1-r^2}}{\sqrt{n-2}} \cdot \frac{s_y}{s_x}}$ ou encore $\boxed{\frac{t_{n-2}}{\sqrt{n-2}} \cdot \sqrt{\frac{1-r^2}{r^2}} \cdot a}$

Le tableau ci-dessous fournit quelques valeurs de t_{n-2} ainsi que du coefficient multiplicatif à appliquer à $\sqrt{\frac{1-r^2}{r^2}} \cdot a$ pour obtenir l'incertitude expérimentale à 95% à partir de n mesures indépendantes.

n	3	4	5	6	7	8	9	10	11	∞
$n-2$	1	2	3	4	5	6	7	8	9	∞
t_{n-2}	12,71	4,30	3,18	2,78	2,57	2,44	2,37	2,31	2,26	1,96
$\frac{t_{n-2}}{\sqrt{n-2}}$	12,71	3,04	1,84	1,39	1,15	1	0,89	0,82	0,75	$\frac{2}{\sqrt{n-2}}$

On obtient une **incertitude relative à 95%** égale à $\boxed{\frac{t_{n-2}}{\sqrt{n-2}} \cdot \sqrt{\frac{1-r^2}{r^2}}}$