Régression linéaire

1 Méthode graphique

1.1 Ajustement à la règle

On trace au jugé une droite $\mathcal{D}$ passant le plus près possible des points du nuage de points, en s'efforçant d'équilibrer le nombre de points situés au dessus et au dessous de la droite $\mathcal{D}$. L'équation de $\mathcal{D}$ est alors de la forme $y = ax + b$. Pour retrouver cette équation, il suffit alors de connaître deux points de $\mathcal{D}$.

1.2 Ajustement affine par la méthode de Mayer

On partage le nuage de points en deux nuages de points de nombres équivalents. On calcule alors le point moyen de chaque nuage qu'on appelle G_1 et G_2 . La droite (G_1G_2) est la droite de Mayer. Elle passe de plus par le centre de gravité du nuage de points noté G. C'est une bonne approximation, si le nuage de points est allongé.

2 Ajustement affine par la méthode des moindres carrés

2.1 Problème

On considère une série statistique à deux variables représentée, dans un repère orthonormé d'origine O, par un nuage de points $M_i(x_i; y_i)$ paraissant justifier un ajustement affine. Le problème est de déterminer quelle droite est susceptible de remplacer "au mieux" ce nuage de points. On souhaite préciser les critères utilisés.

Soit $\mathcal{D}$ une droite d'ajustement. On note P_i le point de même abscisse x_i que M_i situé sur la droite $\mathcal{D}$ d'équation $y = ax + b$. On souhaite que cette droite $\mathcal{D}$ vérifie deux conditions:

• Tout d'abord on souhaite avoir une répartition équilibrée des points m_i du nuage en dessous et au dessus de la droite $\mathcal{D}$. Ce souhait correspond à la relation

$$\sum_{i=1}^{n} \overline{P_iM_i} = 0.$$

• Ensuite on souhaite minimiser la quantité suivante:

$$\sum_{i=1}^{n} \overline{P_iM_i}^2.$$

2.1.1 Première condition

2.1.2 Deuxième condition

On vient de voir que la droite que l'on cherche doit passer par le point moyen G pour vérifier la première condition. Cherchons donc parmi toutes les droites passant par G celle qui minimise la valeur $\sum_{i=1}^{n} \overline{P_iM_i}^2$.

Soit $\mathcal{D}$ une droite d'équation $y = ax + b$ passant par G. Comme G est sur la droite $\mathcal{D}$, on sait que les coordonnées de G vérifient l'équation de la droite, c'est-à-dire: $\overline{y} = a\overline{x} + b$. On en déduite dons la valeur suivante de b:

On veut minimiser la quantité $\sum_{i=1}^{n} \overline{P_iM_i}^2$.

$$\begin{aligned}\sum_{i=1}^{n} \overline{P_iM_i}^2 &= \sum_{i=1}^{n} (y_i - (ax_i + b))^2 \\ &= \sum_{i=1}^{n} (y_i - ax_i - (\overline{y} - a\overline{x}))^2 \\ &= \sum_{i=1}^{n} (y_i - \overline{y} - a(x_i - \overline{x}))^2\end{aligned}$$

Posons $X_i = x_i - \overline{x}$ et $Y_i = y_i - \overline{y}$, on obtient:

$$\begin{aligned}\sum_{i=1}^{n} \overline{P_iM_i}^2 &= \sum_{i=1}^{n} (Y_i - aX_i)^2 \\ &= \sum_{i=1}^{n} (Y_i^2 - 2aX_iY_i + a^2X_i^2) \\ &= \left(\sum_{i=1}^{n} Y_i^2\right) - 2a\left(\sum_{i=1}^{n} X_iY_i\right) + a^2\left(\sum_{i=1}^{n} X_i^2\right)\end{aligned}$$

On cherche donc à minimiser $f(a) = \left(\sum_{i=1}^{n} Y_i^2\right) - 2a\left(\sum_{i=1}^{n} X_i Y_i\right) + a^2\left(\sum_{i=1}^{n} X_i^2\right)$.
On obtient en dérivant:

$$f'(a) = 2a\left(\sum_{i=1}^{n} X_i^2\right) - 2\left(\sum_{i=1}^{n} X_i Y_i\right)$$

. On sait que $f(a)$ atteint son minimum lorsque $f'(a) = 0$, ce qui revient à choisir

$$\begin{aligned} a &= \frac{\sum_{i=1}^{n} X_i Y_i}{\sum_{i=1}^{n} X_i^2} \\ &= \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2} \\ &= \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2} \\ &= \end{aligned}$$

2.1.3 Covariance

Définition 1. La covariance d'une série statistique est définie par:

$$Cov(x, y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}) = \left(\frac{1}{n}\sum_{i=1}^{n} x_i y_i\right) - \overline{xy}.$$

2.1.4 Équations des droites de régression

Quand on cherche une droite de régression de y par rapport à x, les variables y et x ne jouent pas le même rôle:

• y est la variable à expliquer,

• x est la variable potentiellement explicative.

Ceci signifie que l'on a a priori accès à des valeurs "exactes" pour x à partir desquels, on cherche à "prédire" y.
On peut évidemment échanger les rôles de x et y et chercher la droite de régression de x par rapport à y.

Équations des droites de régression:

Définition 2. On appelle droite de régression linéaire par la méthode des moindres carrés de y en x, la droite $\mathcal{D}$ d'équation $y = ax + b$ telle que :
• la droite $\mathcal{D}$ passe par le point moyen G,
• $\sum_{i=1}^{n} \overline{P_i M_i}^2 = \sum_{i=1}^{n}(y_i - (ax_i + b))^2$ est minimale.

Grâce à nos calculs précédents, on peut donner l'équation $y = ax + b$ de la droite $\mathcal{D}$ de régression de y en x où

$$a =$$

et

$$b =$$

soit

$$y =$$

Définition 3. De la même façon, on appelle droite de régression linéaire par la méthode des moindres carrés de x en y, la droite $\mathcal{D}'$ d'équation $x = a'y + b'$ telle que :
• la droite $\mathcal{D}'$ passe par le point moyen G,
• $\sum_{i=1}^{n} \overline{Q_iM_i}^2 = \sum_{i=1}^{n}(x_i - a('y_i + b'))^2$ est minimale.

Avec des calculs similaires, on peut trouver l'équation $x = a'y + b'$ de $\mathcal{D}'$. On obtient

$$a' =$$

et

$$b' =$$

soit

$$x =$$

3 Coefficient de corrélation linéaire

Définition 4. Le coefficient de corrélation linéaire d'une série statistique double de variables x et y est le nombre r défini par:

$$r = \frac{Cov(x,y)}{\sqrt{Var(x)Var(y)}} = \frac{Cov(x,y)}{\sigma(x)\sigma(y)}.$$

Le coefficient de corrélation linéaire est un nombre réel toujours compris entre -1 et 1. C'est un bon indice pour détecter des relations de dépendance entre x et y.

On peut essayer de donner des seuils pour détecter des corrélations fortes ou non mais ces quantités sont arbitraires car elles dépendent fortement de la précision des données fournies.

Interprétation
• Si $r = 1$ ou -1, alors les droites $\mathcal{D}$ et $\mathcal{D}'$ sont confondues et le nuage de points est exactement sur cette droite.
• Si $-1 \leq r \leq \frac{-\sqrt{3}}{2}$ ou $\frac{\sqrt{3}}{2} \leq r \leq 1$, alors il y a une bonne corrélation linéaire, les droites $\mathcal{D}$ et $\mathcal{D}'$ sont presque confondues.
• Si $\frac{-\sqrt{3}}{2} \leq r \leq \frac{1}{2}$ ou $\frac{1}{2} \leq r \leq \frac{\sqrt{3}}{2}$, alors la corrélation linéaire est médiocre et les droites $\mathcal{D}$ et $\mathcal{D}'$ forment un angle important.
• Si $\frac{-1}{2} \leq r \leq \frac{1}{2}$, alors la corrélation est mauvaise et les droites forment presque un angle droit.