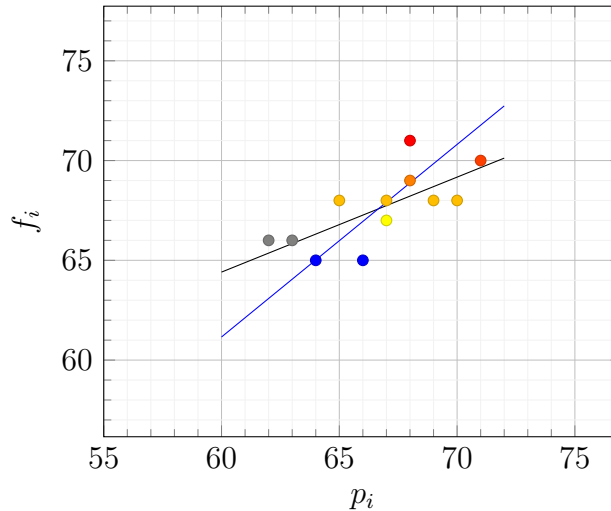


TD I

Exercice 1. L'étude statistique ci-dessous porte sur les poids (en kg) respectifs des pères p_i , et ceux de leurs fils aînés f_i avec $i = 1, \dots, 12$. Les résultats sont tracés sur le graphique suivant



On donne quelques résultats numériques :

$$\sum p_i = 800, \sum p_i^2 = 53418, \sum p_i f_i = 54107, \sum f_i = 811, \sum f_i^2 = 54849$$

1. Calculer la droite des moindres carrés du poids des fils en fonction du poids des pères.
2. Calculer la droite des moindres carrés du poids des pères en fonction du poids des fils.
3. Les deux droites de régression sont-elles identiques?
4. En quel point se coupent ces deux droites? Que vaut le produit des pentes des deux droites?
5. Estimer σ^2 dans le premier modèle de régression. Un père pèse 70 kilos, peut-il raisonnablement espérer que son fils aîné en pèse 80?

Exercice 2. Dans de nombreux cas, lorsque l'on étudie le lien entre Y et X nous savons que si $X = 0$, alors $Y = 0$. On peut alors simplifier le modèle linéaire en cherchant juste à ajuster les points sur une droite d'ordonnée à l'origine nulle. On étudie la régression linéaire $y_i = \beta x_i + \varepsilon_i$, où les ε_i sont centrées, non corrélées et de même variance σ^2 . On considère les deux estimateurs de β suivants

$$\hat{\beta} = \frac{\sum_{i=1}^n x_i y_i}{\sum_{i=1}^n x_i^2} \quad \text{et} \quad \tilde{\beta} = \frac{\sum_{i=1}^n y_i}{\sum_{i=1}^n x_i}$$

1. Quelle est la logique de construction de ces deux estimateurs?
2. Montrer que $\hat{\beta}$ et $\tilde{\beta}$ sont des estimateurs non biaisés de β .
3. Montrer que la variance de $\tilde{\beta}$ est strictement plus grande que la variance de $\hat{\beta}$, sauf dans le cas où les x_i sont tous égaux. Ce résultat était-il prévisible?

Exercice 3. On suppose que le modèle de régression linéaire simple de Y en fonction de X , avec des erreurs centrées et non corrélées, est valide:

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

Montrer qu'alors il en est de même pour le modèle de régression linéaire de X en fonction de Y . Quels sont les paramètres de ce modèle?

Exercice 4. On considère un objet dont le coût de fabrication est x_0 . Supposons que le nombre d'objets vendus en une semaine, y dépend du prix de vente x selon un modèle linéaire simple.

1. Quel est le prix de vente maximisant la marge de l'entreprise?
2. Sur les trois dernières semaines, un industriel a vendu objet

prix	113	115	120
quantité	230	200	125

Sachant que le coût de fabrication est de 100 euros, quel prix de vente lui conseillez-vous?

Exercice 5. Soit y_1, \dots, y_n des réels. Pour mesurer l'écart d'un réel x à l'ensemble des y_i , on peut utiliser la distance $D(x) = \sum_{i=1}^n f(y_i - x)$ où f est une fonction positive, paire, s'annulant en 0, continue et croissante sur les réels positifs.

1. Montrer que le réel qui minimise cette distance lorsque $f(t) = t^2$ est la moyenne des y_i
2. On suppose maintenant que $f(t) = |t|$. Montrer que le réel qui minimise la distance $D(x)$ est la médiane des y_i . *Indication: on considérera que $y_1 \leq \dots \leq y_n$ et on réécrira $D(x)$ sur l'intervalle $[y_j, y_{j+1}[$. On traitera ensuite cas où $n = 2p + 1$, puis le cas où $n = 2p$.*

Exercice 6. Soit y_1, \dots, y_n et x_1, \dots, x_n des réels. Dans le modèle linéaire simple $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, on considère la somme des erreurs en valeur absolue

$$S(\beta_1, \beta_2) = \sum_{i=1}^n |y_i - \beta_0 - \beta_1 x_i|$$

et on notera $\tilde{\beta}_0$ et $\tilde{\beta}_1$ les estimateurs des moindres valeurs absolues, *i.e.* les valeurs de β_0 et β_1 minimisant la fonction $S(\cdot, \cdot)$. On cherche une stratégie pour obtenir $\tilde{\beta}_0$ et $\tilde{\beta}_1$.

1. La valeur de β_1 étant fixée, quelle est la valeur de β_0 qui minimise la fonction $g(\beta_0) = \sum_{i=1}^n |y_i - \beta_0 - \beta_1 x_i|$?
2. En déduire un algorithme pour obtenir $\tilde{\beta}_0$ et $\tilde{\beta}_1$.

Exercice 7. Montrer que le coefficient de détermination R^2 est égal au carré du coefficient de corrélation empirique entre les x_i et les y_i défini par

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

Exercice 8. Dans le modèle linéaire simple, si on considère la normalité des ε_i , on sait que $y_i \sim \mathcal{N}(\beta_0 + \beta_1 x_i, \sigma^2), \forall i = 1, \dots, n$

1. Exprimer la vraisemblance des observations $L(\beta_0, \beta_1, \sigma^2)$.
2. Quelles sont les valeurs de β_0 et β_1 maximisant cette vraisemblance?
3. Quelle est la valeur de σ^2 maximisant cette vraisemblance? Que peut-on alors dire de l'estimateur du maximum de vraisemblance (EMV) de σ^2 ?

Exercice 9. Intervalles de confiance vs Région de confiance On considère le modèle de régression linéaire simple $y = \beta_0 + \beta_1 x + \varepsilon$. Soit un échantillon $(x_i, y_i)_{i=1}^{100}$ de statistiques résumées

$$\sum_{i=1}^{100} x_i = 0 \quad \sum_{i=1}^{100} x_i^2 = 400 \quad \sum_{i=1}^{100} x_i y_i = 100 \quad \sum_{i=1}^{100} y_i = 100 \quad \hat{\sigma}^2 = 1.$$

1. Exprimer les intervalles de confiance à 95% pour $\hat{\beta}_0$ et $\hat{\beta}_1$.
2. Donner l'équation de la région de confiance à 95% de $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1)$. *Rappel : l'ensemble des points (x, y) tels que $\frac{(x-x_0)^2}{a^2} + \frac{(y-y_0)^2}{b^2} \leq 1$ est l'intérieur d'une ellipse centrée en (x_0, y_0) dont les axes sont parallèles à ceux des abscisses et des ordonnées, et de sommets $(x_0 \pm a, 0)$ et $(0, y_0 \pm b)$.*
3. Représenter sur un même graphique les résultats obtenus.