# TD 1 : Régression linéaire simple

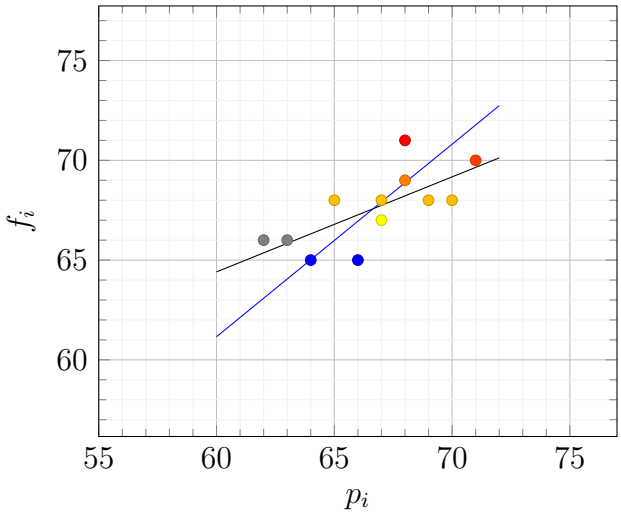
**Exercice 1. Rappels de cours**

1. Rappeler le principe d'une régression linéaire simple. Préciser les hypothèses.
2. Rappeler les définitions des quantités suivantes : $x_i$, $y_i$, $\beta_0$, $\beta_1$, $\bar{x}$, $\bar{y}$, $\hat{\beta}_0$, $\hat{\beta}_1$, $\epsilon_i$, $\hat{\epsilon}_i$. On indiquera en particulier si ces quantités sont déterministes ou aléatoires.
3. Faire un schéma pour donner une interprétation géométrique à la régression linéaire simple.
4. Donner la définition du coefficient de détermination $R^2$ et son interprétation. Montrer que ce coefficient s'écrit

$$R^2 = \frac{[\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})]^2}{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2} =: \frac{s_{xy}^4}{s_x^2 s_y^2} = \rho(x, y)^2 .$$

   où $\rho(x, y)$ est le coefficient de corrélation empirique entre $x$ et $y$.
5. Donner les hypothèses supplémentaires dans le cas d'une régression linéaire gaussienne. Quelles informations a-t-on en plus dans ce contexte ?

**Exercice 2.** 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 $i = 1, \cdots, 12$. 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 ? Identifiez-les sur le graphique.
4. En quel point se coupent ces deux droites ? Que vaut le produit des pentes des deux droites ?
5. Estimer $\sigma^2$ dans le premier modèle de régression. Un père pèse 70 kilos, peut-il raisonnablement espérer que son fils aîné en pèse 80 ?

**Exercice 3.** 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 $\varepsilon_i$ sont centrées, non corrélées et de même variance $\sigma^2$. On considère les deux estimateurs de $\beta$ 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 $\beta$.
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. (On pourra utiliser l'inégalité de Cauchy-Schwarz.) Ce résultat était-il prévisible ?

**Exercice 4.** 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 5.** On considère un produit dont le coût de fabrication est $x_0$. Supposons que le nombre de produits 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 fait varier le prix du produit. On dispose des données suivantes.

| | | | |
|---|---|---|---|
| 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 6.*** Soit $y_1, \ldots, y_n$ 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 : considérer que $y_1 \leq \cdots \leq y_n$, réécrire $D(x)$ sur l'intervalle $[y_j, y_{j+1}[$, puis tracer $D(x)$. On traitera le cas où $n = 2p+1$, puis le cas où $n = 2p$.*

**Exercice 7.*** Soit $y_1, \ldots, y_n$ et $x_1, \ldots, x_n$ 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 $\beta_0$ et $\beta_1$ minimisant la fonction $S(\cdot,\cdot)$. On cherche une stratégie pour obtenir $\tilde{\beta}_0$ et $\tilde{\beta}_1$.

1. La valeur de $\beta_1$ étant fixée, quelle est la valeur de $\beta_0$ 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 8. Retour sur le cas gaussien**

Dans le modèle linéaire simple, si on considère la normalité des $\varepsilon_i$, alors

$$y_i \sim \mathcal{N}\left(\beta_0 + \beta_1 x_i, \sigma^2\right), \quad i = 1, \ldots, n.$$

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

**Exercice 9. Fréquence Cardiaque** On s'intéresse à la fréquence cardiaque de sportifs amateurs mesurée après trois quarts d'heure d'un effort soutenu. On veut déterminer si l'âge du sportif a une influence sur sa fréquence cardiaque après un effort soutenu. On dispose de $n = 40$ observations du couple $(y_i; x_i)$ où $y_i$ est la fréquence cardiaque et $x_i$ l'âge du sportif $i$. On suppose un modèle linéaire gaussien classique entre $y$ et $x$, d'intercept $\beta_0$ et de pente $\beta_1$. On donne :

$$\bar{y} = 171.3 \; ; \; \bar{x} = 38.4 \; ; \; \sum_{i=1}^{40}(x_i - \bar{x})^2 = 4381 \; ; \; \sum_{i=1}^{40}(y_i - \bar{y})^2 = 424 \; ; \; \sum_{i=1}^{40}(x_i - \bar{x})(y_i - \bar{y}) = -961$$

1. Donnez l'expression de $\hat{\beta}_0$ et $\hat{\beta}_1$, et calculez leur valeur.
2. Calculez le coefficient de détermination $R^2$. Interprettez.
3. Donnez la loi de $\hat{\beta}_1$ en fonction de $\sigma^2$.
4. Donnez la définition de $\hat{y}_i$ et $\hat{\epsilon}_i$. Que représente cette quantité ? On donne :

$$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = 213.3.$$

   En déduire un estimateur $\hat{\sigma}^2$ sans bias pour la variance, puis un estimateur $\hat{\sigma}_1^2$ pour la variance de $\hat{\beta}_1$.
5. Testez l'hypothèse $H_0 : \beta_1 = 0$ v.s. $H_1 : \beta_1 \neq 0$. On donne le quantile de la loi de Student à 38 degrés de libertés à 97.5%: $q_{t,38}(97.5\%) = 2.024394$. Interpettez.
6. Pouvez-vous prédire la fréquence cardiaque d'un sportif de 57 ans ? De 0 ans ? De 817 ans ? Donnez les intervalles de prédictions. Interpettez.

**Exercice 10. 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 $\hat{\beta}_1$. On admet que le quantile d'ordre 0.975 d'une loi de Student à 98 degrés de liberté vaut environ 2.
2. Donner l'équation de la région de confiance à 95% de $\hat{\beta} = \left(\hat{\beta}_0, \hat{\beta}_1\right)$. On admet que le quantile d'ordre 0.975 d'une loi de Fisher à $(2, 100)$ degrés de liberté 3.
   *Rappelons que 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.