Université Joseph Fourier
L2/STA230

TP : Regression linéaire

L'objectif de ce TP est de savoir construire un **modèle de regression linéaire** visant à expliquer une variable par une autre. Nous nous intéressons ici au poids des bébés et à l'âge de leur mère dans la population des bébés issus de mère non fumeuse et d'âge compris entre 21 et 35 ans. Vous allez construire un modèle linéaire cherchant à expliquer linéairement le poids des bébés par l'âge de la mère, prédire le poids de bébés inconnus à partir de l'âge supposé de sa mère et enfin réaliser un test statistique pour répondre à la question "Peut on expliquer le poids des bébés par l'âge de la mère ?"
Vous disposez également du document "regression mères fumeuses.ods",

1. Ouvrez le fichier de données "BB mère non fumeuses.csv".
2. Calculez les caractéristiques usuelles : nombre d'observations, âge moyen, poids moyen, variance des âges, variance des poids, covariance.
3. Déterminez la droite de régression, $y = \hat{a}x + \hat{b}$ à partir des résultats précédents et confirmez vos résultats à l'aide des fonctions adaptées d'Open Office.
4. Déterminez le coefficient de corrélation linéaire empirique r, ainsi que le coefficient de détermination r^2, en fonction des caractéristiques et à l'aide des fonctions adaptées d'OpenOffice Calc. Commentez ces résultats.
5. Faites la représentation graphique des données et superposez au graphique la droite de regression.
6. Déterminez les valeurs ajustées, $\hat{y}_i = \hat{a}x_i + \hat{b}$, ainsi que les résidus correspondants $\hat{\epsilon}_i = y_i - \hat{y}_i$, pour $i = 1, \ldots, n$. Retrouvez les valeurs ajustées avec la fonction adaptée d'OpenOffice Calc.
7. Quelle est la prédiction du poids d'un bébé dont la mère est agée de 43 ans ?
8. Représentez les résidus en fonction des valeurs ajustées. Les résidus ont ils une structure particulière ?
9. On souhaite répondre à la question "Le modèle linéaire permet-il d'expliquer le poids des bébés par l'âge de la mère ?". Cette question peut se traduire par "le coefficient directeur de la droite (a) est il différent de 0 ?". Pour y répondre, vous allez calculer la version studentisée du coefficient a, a^s_{obs}. La loi du coefficient studentisée étant connue, il vous est possible en construisant un test statistique de déterminer si a^s_{obs} est significativement loin de 0. Calculez la p valeur associée au test et interprétez le résultat.
10. Quel lien faîtes vous entre ce résultat et le résultat du test d'indépendance entre l'âge et le poids, réalisé lors du TP précédent ?
11. De manière similaire, construisez le test statistique permettant de comparer l'ordonnée (b) à l'origine à 0.