TP de L2 SPI parcours informatique

Outils statistiques pour l'informatique

Régression linéaire.

Une régression est un modèle qui permet de prédire la valeur d'une variable Y en fonction de la valeur d'une autre variable X. Cela n'est possible que s'il existe une dépendance entre Y et X : on parle alors de corrélation.

Une régression linéaire simple est une relation affine entre Y et une seule variable X, de la forme :

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

Avec :

- Y est la la variable dépendante
- β_0 représente la valeur moyenne des observations Y_i quand $x_i = 0$ (=ordonnée à l'origine)
- β_1 représente la pente de la droite de régression, et correspond à la variation moyenne de Y si X augmentait d'une unité
- X est la variable explicative.
- ε est une erreur aléatoire, que nous ne considérerons pas dans cet exercice.

Pour calculer β_0 et β_1, il faut appliquer les formules suivantes :

$$\beta_1 = \frac{\Sigma(X_i - \bar{X})(Y_i - \bar{Y})}{\Sigma(X_i - \bar{X})^2}$$

$$\beta_0 = \bar{Y} - \beta_1 \bar{X}$$

Soient les données présentées dans le tableau ci-dessous. Il s'agit du nombre de calories consommées par jour et du pourcentage de population agricole dans 11 pays.

Observation i	Pays	% Population agricole	Calories par jour et par personne
1	Suisse	4,0	3 432
2	France	5,7	3 273
3	Suède	4,9	3 049
4	USA	3,0	3 642
5	Ex-URSS	14,8	3 394
6	Chine	69,6	2 628
7	Inde	63,8	2 204
8	Brésil	26,2	2 643
9	Pérou	38,3	2 192
10	Algérie	24,7	2 687
11	Ex-Zaire	65,7	2 159

Les exercices suivants doivent être réalisés en Python.

1. Représenter graphiquement Y (calories consommées par jour) en fonction de X (pourcentage de population agricole).

2. Estimer les paramètres $\beta_0 + \beta_1$ du modèle :

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

3. Représenter la droite de régression sur le graphique de la question 1.

4. Calculer la valeur du coefficient de corrélation de Pearson ρ entre X et Y, à partir de l'estimateur r définit tel que :

   $$r = \frac{\Sigma(X-\bar{X})(Y-\bar{Y})}{\sqrt{\Sigma(X-\bar{X})^2}\sqrt{\Sigma(Y-\bar{Y})^2}}$$

   (la valeur de ρ est donnée par r)