

## Intelligence Artificielle et Données : feuille 2

### Analyse de données Régression linéaire simple et multiple

## 1 Régression simple

### Exercice 1.

La gérante d'un commerce veut évaluer l'impact des frais déboursés en publicité par mois (représentés par une variable  $X$  exprimée en milliers d'euros) sur le chiffre d'affaires mensuel (représenté par une variable  $Y$  exprimée en milliers d'euros). On aimerait évaluer dans quelle mesure une modification du budget publicitaire mensuel affecterait le chiffre d'affaires mensuel. On a donc recueilli sur une période de 10 mois les données du tableau ci-dessous.

chiffre d'affaires	220	280	250	170	150	340	310	210	180	190
frais publicitaires	2.6	2.6	2.4	1.5	0.9	3.0	2.7	2.3	1.7	1.9

En vous aidant des bonnes bibliothèques python, répondez aux questions suivantes :

1. Tracer le nuage de points et estimer le coefficient de corrélation linéaire.
2. Etablir la droite de régression correspondant à ce problème et tracer cette droite.
3. Tester la significativité de la régression au risque 5%.
4. Calculer le coefficient de détermination.
5. Quel serait le chiffre d'affaires mensuel prédit par le modèle pour un budget publicitaire mensuel de 400 euros ? de 4000 euros ?

### Exercice 2. (Travail à rendre)

La société Métalex moule des pièces dans un four. L'ingénieur se demande s'il existe un lien entre la température (en degré celsius) à laquelle les pièces sont moulées et leur résistance (en  $\text{kg}/\text{cm}^2$ ). Il dispose des données suivantes transmises par l'atelier.

température	100	120	140	160	180	200	220	240	260	280	300
résistance	46	48	49	51	52	53	54	55	56	56	56

1. Tracer le nuage de points et estimer le coefficient de corrélation linéaire.
2. Ajuster un modèle linéaire de la forme  $Y = aX + b + \varepsilon$  : établir la droite de régression correspondant à ce problème et tracer cette droite.
3. Déterminer un intervalle de confiance au risque 5% des paramètres de la droite de régression.
4. Tester la significativité de la régression au risque 5%.
5. Calculer le coefficient de détermination.

## 2 Régression multiple

### Exercice 1.

Nous sommes intéressés à savoir quels sont les facteurs importants qui influencent ou déterminent la valeur d'une propriété et de construire un modèle qui nous aidera à évaluer cette valeur selon ces facteurs. Pour ce faire, nous avons obtenu la valeur totale pour un échantillon de 79 propriétés dans une région donnée. Voir le fichier `proprietes.xls` disponible à l'adresse :

[http://www.labri.fr/~zemmari/ia\\_data/datasets](http://www.labri.fr/~zemmari/ia_data/datasets)

Le fichier contient les variables suivantes correspondant chacune à un facteur :

- Valeur Totale : valeur globale de la propriété
- Valeur Terrain : valeur du terrain sur lequel est bâtie la maison
- Acre : Superficie en acres
- Pieds2 : Superficie en  $\text{pieds}^2$
- Extérieur : Etat extérieur
- Chauffage : Type du chauffage utilisé
- Pièces : Nombre de pièces dans la maison
- Chambres : Nombre de chambres dans la maison
- SbainsC : Nombre de salles de bain complètes
- Sbains : Nombre de salles de bain complètes
- Foyers : Nombre de foyers
- Garage : Indique si la propriété contient un garage ou non.

1. Y a-t-il un lien entre la valeur totale et les différents facteurs ?
2. Afficher les valeurs des différentes statistiques descriptives.
3. Faites les différentes régressions linéaires simples possibles, et choisissez la (ou les) plus pertinente(s).
4. Faites la régression multiple de la Valeur totale en fonction des autres facteurs (quand c'est possible). Analysez les résultats obtenus.