Regression linéaire

Exercice 0

Le service des études économiques d'une société cherche à mesurer l'incidence de la modulation de la pression marketting (variable explicative) sur la vente de boites de conserve (variable à expliquer). Il procède à une expérience dans cinq zones géographiques de caractéristiques voisines et enregistre les ventes (en milliers de boites) réalisées dans chaque zone durant une même période, ainsi que les dépenses (en milliers de dollards) consenties par la firme pour les budgets de publicité (locale) et de promotion des ventes. Les résultats sont répertoriés dans le tableau suivant :

ventes	25	30	35	45	65
depenses	5	6	9	12	18

1. Définir les vecteurs ventes et dépenses représentant ces données. Identifier la variable réponse et la variable explicative.
2. Tracer le nuage de points. Commenter

Exercice 1

On dispose de données concernant l'âge (age), le kilométrage en milliers de kilomètres (km) et le prix en milliers d'euros (prix) pour un échantillon de voitures d'occasion du même type.

age	5	4	6	5	5	5	6	6	7	7
km	92	64	124	97	79	76	93	63	111	143
prix	7.8	9.5	6.4	7.5	8.1	9	6.1	9.7	6.4	4.4

On s'intéresse aux couples de variables (age, prix) et (km, prix). Le but étant, si le lien est fort, de modéliser la variable prix en fonction d'une des deux variables explicatives km ou age.

1. Quelle est la première chose à effectuer avant tout calcul ? Effectuez-la et commenter le résultat obtenu.
2. Quelle est la deuxième étape à franchir ?
3. En fonction des résultats précédemment obtenus, choisir la variable explicative à garder puis :

(a) Effectuer la régression linéaire sous R. Tracer la droite de régression sur le nuage de points.

(b) Donner l'équation de la droite de régression et interpréter ses coefficients.

(c) Vérifier numériquement et graphiquement que la droite de régression passe par le point central des données (dit centre de gravité).

(d) Peut-on s'appuyer sur le modèle pour faire de la prédiction ? Si oui, determiner le prix que l'on peut prédire pour une voiture dont le kilométrage est de 150.000 km.

Exercice 2

On cherche à étudier l'évolution du montant des achats réalisés en ligne par les ménages français. Le tableau ci-dessous donne une estimation du montant de ces achats sur sept années consécutives, de 1998 à 2004, codées par les rangs 0 à 6 :

Rang de l'année	0	1	2	3	4	5	6
Montant d'achats en millions d'euros	75	260	820	1650	2300	4000	5300

1. (a) Pour répondre au but de l'étude, quelle variable devons-nous identifier comme variable explicative ? Comme variable à expliquer ?

(b) Tracer le nuage de points. Commenter.

(c) Donner le coefficient de corrélation linéaire entre x et y. Interpréter le résultat obtenu.

(d) Donner l'équation de la droite de régression de y en x

(e) Tracer la droite de régression (en bleu) sur le nuage de points.

(f) Quelle prévision du montant d'achats peut-on faire pour l'année 2005? Est-elle fiable?

2. On considère la nouvelle variable réponse $z = \sqrt{y}$.

(a) Répondre aux questions b) à f) de la question 1.

(b) En déduire une expression de y en fonction de x, puis une prévision du montant d'achats pour l'année 2005.

3. À partir du tableau de données, un logiciel statistique propose un ajustement polynomial par l'équation y = 130x2 + 100x + 68. Déduire de cet ajustement une prévision du montant d'achats pour l'année 2005.

4. Le montant des achats en ligne en 2005 a été de 7700 millions d'euros. Lequel des ajustements précédents a été le plus conforme à la réalité ? Pourquoi ?

Exercice 3

Une étude a été menée sur les caractéristiques de 31 arbres de la même espèce (cerisier tardif). Pour chacun de ces arbres, on dispose de mesures de leur diamètre et de leur hauteur en pieds (un pied équivaut à à peu près un tiers de mètre). On cherche à étudier la manière dont la hauteur des arbres dépend de leur diamètre.

Diam	8.3	8.6	8.8	10.5	10.7	10.8	11.0	11.0	11.1	11.2	11.3	11.4	11.4	11.7	12.0	12.9
Haut	70	65	63	72	81	83	66	75	80	75	79	76	76	69	75	74
Diam	12.9	13.3	13.7	13.8	14.0	14.2	14.5	16.0	16.3	17.3	17.5	17.9	18.0	18.0	20.6	
Haut	85	86	71	64	78	80	74	72	77	81	82	80	80	80	87	

1. Quelle est la variable réponse ? La variable explicative ?

2. Représentez les données sous la forme d'un nuage de points. Qu'observez-vous sur ce graphique ?

3. Calculer le coefficient de corrélation entre les variables. Que pouvez-vous conclure sur la liaison entre le diamètre des arbres et leur hauteur ? Cela confirme-t-il les observations faites sur le nuage de points ? Pensez-vous qu'une modélisation linéaire soit adaptée ?

4. Imaginons que l'on observe en plus un arbre "hors norme" de diamètre 50 pieds et de hauteur 114,75 pieds. Calculez le nouveau coefficient de corrélation. Pensez-vous que la modélisation linéaire soit pour autant plus adaptée? Conclure sur l'utilité de visualiser le nuage de points et les problèmes que peuvent poser les valeurs extrêmes.