TP2 : Régression linéaire

• Exercice 1

Une entreprise fixe des prix différents pour un produit particulier dans huit régions différentes des États-Unis. Elle souhaite étudier la liaison éventuelle entre le nombre de ventes (variable Y) et le prix du produit (variable X). Pour les n = 8 régions, on observe les valeurs $(x_1, y_1), \dots, (x_n, y_n)$ de (X,Y) suivantes :

x_i	420	380	350	400	440	380	450	420
y_i	5.5	6	6.5	6	5	6.5	4.5	5

1. Représenter le nuage de points $\{(x_1, y_1), \dots, (x_n, y_n)\}$. À partir de celui-ci, expliquer pourquoi on peut envisager l'existence d'une liaison linéaire entre Y et X.
2. On adopte alors le modèle de régression linéaire simple : $\forall i, y_i = \beta_0 + \beta_1 x_i + \in_i$. Les paramètres β_0 et β_1 sont des réels inconnus. On considère la forme matricielle usuelle : Y = Xβ + ϵ , avec $\beta = \begin{pmatrix}\beta_0 \\ \beta_1\end{pmatrix}$. Créer dans R la matrice X associée.
3. En posant y = ${}^t(y_1, \dots, y_n)$, calculer $b = \left({}^tXX\right)^{-1}\ {}^tXy$. Que représente b par rapport à β ?
4. Vérifier que l'on a b = ^{t}(b0 ,b1) , avec

   $b1 = \frac{1}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$

   $b0 = \bar{y} - b1\bar{x}$.

   Retrouver ces résultats numériques avec les commandes lm et coef.
5. Tracer la droite de régression sur le nuage de points.
6. Calculer "à la main" le coefficient de détermination et le coefficient de détermination ajusté. Est-ce que le modèle de régression linéaire simple est pertinent avec les données ?
7. Retrouver les estimations précédentes avec la commande summary.

• Exercice 2

L'entreprise CITRON fabrique un matériau en matière plastique qui est utilisé dans la fabrication de jouets. Le département de contrôle de qualité de l'entreprise a effectué une étude qui a pour but d'établir dans quelle mesure la résistance à la rupture (en kg/cm2) de cette matière plastique pouvait être affectée par l'épaisseur du matériau ainsi que la densité de ce matériau. Douze essais ont été effectués et les résultats sont présentés dans le tableau ci-dessous.

Essai numéro	Résistance à la rupture (Y)	Epaisseur du matériau (X1)	Densité (X2)
1	37.8	4	4
2	22.5	4	3.6
3	17.1	3	3.1
4	10.8	2	3.2
5	7.2	1	3.0
6	42.3	6	3.8
7	30.2	4	3.8
8	19.4	4	2.9
9	14.8	1	3.8
10	9.5	1	2.8
11	32.4	3	3.4
12	21.6	4	2.8

1. Compléter les tableaux suivants :

   Régression de la résistance à la rupture Y en fonction de l'épaisseur X_1

Coefficients ($\hat{\beta}_j$)	Erreurs-types s($\hat{\beta}_j$)
b0 = 3.5226	4.383
b1 = 6.03591	1.279

Source de variation	Somme des carrés	ddl = Df = degre de liberté = degree freedom
Régression X1	980.63	1
Résiduelle	440.03	10

2. Compléter les tableaux suivants :

   Régression de la résistance à la rupture Y en fonction de la densité X_2

Coefficients ($\hat{\beta}_j$)	Erreurs-types s($\hat{\beta}_j$)
b0 = -36.373	20.489
b1 =17.4644	6.069

Source de variation	Somme des carrés	ddl
Régression X2	643.57	1
Résiduelle	777.10	10

3. Compléter les tableaux suivants :

   Régression de la résistance à la rupture Y en fonction de l'épaisseur X_1 et de la densité X_2

Coefficients ($\hat{\beta}_j$)	Erreurs-types s($\hat{\beta}_j$)
-30.081	11.455
4.905	1.014
11.072	3.621

Source de variation	Somme des carrés	ddl
Régression (X1,X2)	980.63 / 224.22	1
Résiduelle	215.81	9

4. Quel pourcentage de variation dans la résistance à la rupture est expliquée par chacune des régressions ? somme carré de la regression de x1 / somme c r x1 + residu
5. Compléter le tableau suivant :

	Carré moyen résiduel	Ecart-type des résidus
Régression due à X1	44.00	6.633
Régression due à X2	77.71	8.815
Régression due à (X1,X2)	23.98	4.897

6. Compléter le tableau d'analyse de la variance suivant pour la régression comportant les deux variables explicatives.

Source de variation	Somme des carrés	Ddl	Carrés moyens	Fobs
Régression due à (X1,X2)				
Résiduelle				
Totale				

7. Tester au seuil de signification 5%, l'hypothèse nulle H0 : $\beta 1 = \beta 2 = 0$ contre l'hypothèse alternative H1 : au moins un des coefficients est différent de 0. Quelle est votre conclusion ?
8. Dans le cas du modèle de régression linéaire ne comportant que l'épaisseur du matériau comme variable explicative, déterminer un intervalle de confiance à 95% pour $\beta 1$. Pouvons-nous affirmer, au seuil de signification 5%, que la régression linéaire est significative entre la résistance à la rupture et l'épaisseur du matériau ? Justifier votre conclusion.
9. Quel est l'apport marginal de la variable X_2 lorsqu'elle est introduite à la suite de la variable X_1 ?
10. Nous voulons obtenir diverses estimations et prévisions de la résistance à la rupture. Quelle est, en moyenne, la résistance à la rupture de jouets dont l'épaisseur du matériau utilisé et la densité du matériau sont celles indiquées dans le tableau suivant ?

Epaisseur X1	Densité X2
4	3.8
3	3.4
4	2.9