Feuille de TD n°6

Exercice n°1 :
Un bureau de conseil en ressources humaines a effectué une étude sur le niveau d'anxiété Y mesuré sur une échelle de 1 à 50 de cadres d'entreprises au cours d'une période de deux semaines. On souhaite examiner si les facteurs suivants peuvent avoir une influence sur le niveau d'anxiété des cadres :

- X_1 : pression artérielle systolique
- X_2 : test évaluant les capacités managériales
- X_3 : niveau de satisfaction du poste occupé.

Le tableau d'analyse de la variance indique l'apport de chaque variable introduite dans l'ordre indiqué et ceci pour 22 cadres.

Source de variation	Somme des carrés	ddl
Régression due à X_1	981,326	1
Régression due à X_2	190,232	1
Régression due à X_3	129,431	1
Résiduelle	442,292	18
Totale	1743,281	21

1. Quelle est la somme des carrés dues à la régression pour l'ensemble des trois variables explicatives?
2. Quelle proportion de la variation dans le niveau d'anxiété est expliquée par les trois variables explicatives?
3. Peut-on conclure que dans l'ensemble, les trois variables explicatives ont un effet significatif sur le niveau d'anxiété? Utiliser un seuil de signification de $\alpha = 5\%$. Préciser les hypothèses que l'on désire tester.
4. Si nous ne tenons compte dans le modèle que de la variable X_1, quel est alors le tableau de variance associé?
5. Tester les hypothèses nulles suivantes, au seuil de signification $\alpha = 5\%$, en utilisant un test F approprié :
   (a) $\mathcal{H}_0$: $\beta_1 = 0$ dans le modèle $Y = \beta_0 + \beta_1 X_1 + \epsilon$;
   (b) $\mathcal{H}_0$: $\beta_2 = 0$ dans le modèle $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$;
   (c) $\mathcal{H}_0$: $\beta_3 = 0$ dans le modèle $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \epsilon$;

Exercice n°2 :
On souhaite étudier la variation du taux d'hémoglobine dans le sang au cours d'une opération chirurgicale en fonction de la durée de cette opération et du volume de sang perdu pendant celle-ci. On dispose de $n = 8$ observations. Pour $i \in \{1, \dots, 8\}$, y_i représente la valeur obsservée (en pourcentage) de la variation du taux d'hémoglobine, x_{i1} la durée de l'opération (en heures) et x_{i2} le volume de sang perdu (en litre).

y_i	-1.7	-4.61	-5.82	-1.17	-4.23	-3.31	0.42	-2.98
x_{i1}	1.75	1.33	1.43	1.86	1.81	1.66	1.6	2
x_{i2}	0.52	0.59	0.61	0.5	0.54	0.49	0.27	0.47

On note X la matrice d'expérience : $X = [1|X_1|X_2]$.

1. Après calcul, on a obtenu les résultats suivants :

$$\bar{x_1} = 13.44 \quad \bar{x_2} = 3.99 \quad \sum(x_{i1}^2) = 22.93 \quad \sum(x_{i2}^2) = 2.07 \quad \sum(x_{i1})(x_{i2}) = 6.66$$

   Déterminer alors la matrice $X'X$.
2. On a également trouvé : $Y'X = (-23.4, -38.05, -12.93)$ Déterminer alors des estimations des paramètres du modèle de régression linéaire multiple.

Exercice n°3 :
On considère un modèle de régression $E(Y) = X\beta$ et $var(Y) = \sigma^2 W^{-1}$ avec W une matrice symétrique définie positive connue et σ^2 inconnue.

1. Montrer que $W^{1/2}$ existe et donner son expression.
2. Exprimer le modèle de régression en fonction de $\tilde{Y} = W^{1/2}Y$. Décrire ce nouveau modèle.
3. En déduire un estimateur de β et donner des propriétés. Dans le cas d'un modèle gaussien, donner sa loi.
4. On considère le modèle avec données répétées suivant :

$$\forall i \in \{1, \dots n\}, \quad \forall j \in \{1, \dots, m_i\}, \quad E(Y_{ij}) = a + bx_i \quad et \quad var(Y_{ij}) = \sigma^2,$$

   les observations étant supposées être indépendantes les unes des autres.
   En fait, on suppose que seulement les réponses moyennes par groupe sont observées. On notera y_i la réponse moyenne pour le i-ème groupe :

$$y_i = \frac{1}{m_i} \sum_{j=1} m_i y_{ij}.$$

   Ecrire me modèle correspondant aux observations et donner une expression pour les estimations des paramètres de ce modèle.
   Application numérique : on onserve $n = 4$ valeurs différentes de x chacune 3 fois : $m_1 = m_2 = m_3 = m_4 = 3$. On a obtenu les résultats suivants :

$x_1 = 1.0$	$y_{11} = 2.977$ $y_{12} = 3.009$ $y_{13} = 3.128$	$x_1 = 1.5$	$y_{11} = 3.195$ $y_{12} = 3.138$ $y_{13} = 2.987$
$x_1 = 2.0$	$y_{11} = 3.221$ $y_{12} = 3.192$ $y_{13} = 3.272$	$x_1 = 2.1$	$y_{11} = 3.044$ $y_{12} = 3.246$ $y_{13} = 3.155$

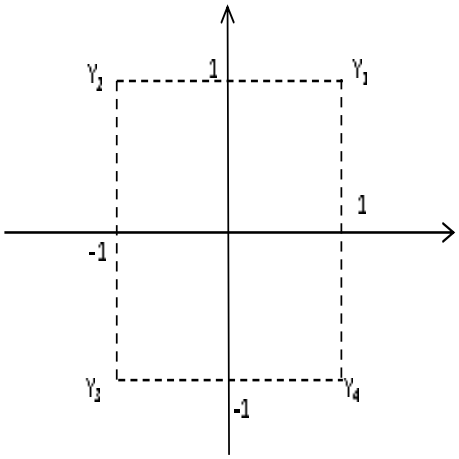
Donner une estimation des paramètres a et b.

Exercice n°4 :
On considère l'effet de deux facteurs quantitatifs, la température et la durée d'une réaction chimique, sur le rendement de cette réaction. On utilise le modèle de régression suivant :

$$E(Y_i) = \beta_0 + \beta_1 t_i + \beta_2 d_i$$

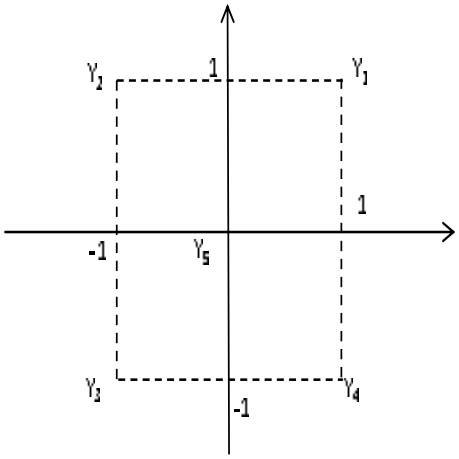
où t_i et d_i représentent respectivement la température et la durée de l'expérience i. Les valeurs possibles pour la température ont 60, 90 et 120 degrés et pour la durée 20, 40 et 60 minutes. Commen les valeurs pour les deux facteurs sont espacées régulièrement, on peut alors procéder au codage suivant : $\{-1, 0, 1\}$. On parle de modèle sans intéraction car on ne considère pas de terme croisé entre la température et la durée d'une réaction chimique dans le modèle ci-dessus.

1. On procède à un premier plan d'expérience décrit par la figuree ci-dessous :



   (a) Ecrire la matrice X de plan d'expérience.
   (b) Calculer son rang.
   (c) On observe les valeurs suivantes : $y_1 = 6.702$, $y_2 = 2.698$, $y_3 = 1.846$ et $y_4 = 5.578$. Donner une estimation des paramètres du modèle.

2. On ajoute une expérience Y_5. La figure ci-dessous représente le nouveau plan d'expérience.



   Ecrire la nouvelle matrice X et calculer son rang. A l'aide de l'observation $y_5 = 4.661$ donner une nouvelle estimation des paramètres.

Remarque 1
Un petit rappel sur les tests entre modèles emboîtés.
Soit le modèle et les hypothèses :

$$Y = X\beta + \epsilon \quad o \quad \epsilon \sim \mathcal{N}(0, \sigma^2 I)$$

On souhaite tester la nullité simultanée des q derniers coefficients du modèle avec $q < p$ (p étant le nombre de paramètres du modèle).
Le problème s'écrit de façon équivalente :

$$\mathcal{H}_0 \ : \quad \beta_{p-q+1} = \ldots = \beta_p = 0$$

contre

$$\mathcal{H}_1 \ : \quad \exists j \in \{p+q-1, \ldots p\}, \ \ \beta_j \neq 0$$

Sous l'hypothèse $\mathcal{H}_0$, le modèle devient :

$$Y = X_0\beta_0 + \epsilon_0 \quad o \quad \epsilon_0 \sim \mathcal{N}(0, \sigma^2 I)$$

où X_0 est composée des $p - q$ premières colonnes de X.
Pour tester ces deux hypothèses, nous utilisons alors la statistique de test F ci-dessous qui possède comme loi sous $\mathcal{H}_0$:

$$F = \frac{||\hat{Y}_0 - \hat{Y}||^2/(p - p_0)}{||Y - \hat{Y}||^2/(n - p)} \sim \mathcal{F}_{p-p_0, n_p}$$

Ici p_0 représente le nombre de colonne de X_0.

Remarque 2
Une table d'analyse de la variance est de la forme, par exemple :

Source de variation	*Somme des carrés*	*DDl*	*carrés moyens*
Expliquée	$SCE = \sum(\hat{y}_i - \bar{y})^2$	p	$CME = \frac{SCE}{p}$
Résiduelle	$SCR = \sum(y_i - \hat{y}_i)^2$	$n - p - 1$	$CMR = \frac{SCR}{n-p-1}$
Totale	$SCT = \sum(y_i - \bar{y}_i)^2$	$n - 1$	