Université de Lorraine - Site de Metz - UFR MIM **Année 2012/2013**
Master de Mathématiques, année M1
Module *Introduction à l'analyse des données*

Exercices - Feuille 4

Régression linéaire multivariée

1- Quelques identités utiles

1) Pour un échantillon $(x_i, y_i)_{1\leq i\leq n}$, on introduit les quantités suivantes

$$S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2, \quad S_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}), \quad S_{yy} = \sum_{i=1}^{n}(y_i - \bar{y})^2. \tag{1}$$

Vérifier les formules pratiques suivantes pour calculer S_{xx}, S_{xy}, S_{yy} sont

$$S_{xx} = \sum_{i=1}^{n} x_i^2 - \frac{1}{n}(\sum_{i=1}^{n} x_i)^2, \quad S_{xy} = \sum_{i=1}^{n} x_i y_i - \frac{1}{n}(\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i), \quad S_{yy} = \sum_{i=1}^{n} y_i^2 - \frac{1}{n}(\sum_{i=1}^{n} y_i)^2. \tag{2}$$

2) On note $\hat{y}_i$ la valeur prédite par la droite de régression. On définit alors

- Somme totale des carrés
$$SCT = \sum_{i=1}^{n}(y_i - \bar{y})^2 = S_{yy} \tag{3}$$
- Somme des carrés de régression
$$SCR = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 \tag{4}$$
- Somme des carrés des erreurs
$$SCE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{5}$$

Vérifier que

$$SCT = SCR + SCE \tag{6}$$

Vérifier aussi que

$$SCR = \frac{S_{xy}^2}{S_{xx}}, \quad SCE = S_{yy} - \frac{S_{xy}^2}{S_{xx}} \tag{7}$$

En déduire les formules pratiques pour évaluer rapidement SCR, SCR, SCE.

3) Rappeler la définition du coefficient de coorélation de Bravais-Pearson r_{xy}. On introduit également le coefficient de détermination R défini par

$$R^2 = \frac{SCR}{SCT} \tag{8}$$

Vérifier que $R^2 = r_{xy}^2$, c-a-d., $R = |r_{xy}|$.
4) On rappelle que la droite de régression empirique est donné par $y = \hat{\beta}_0 + \hat{\beta}_1 x$, avec

$$\beta_1 = \frac{S_{xy}}{S_{xx}} \tag{9}$$

et $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$. Montrer que

$$\hat{\beta}_1 = \sum b_i y_i \tag{10}$$

avec

$$b_i = \frac{x_i - \bar{x}}{S_{xx}} \tag{11}$$

et que

$$\hat{\beta}_0 = \sum a_i y_i \tag{12}$$

avec

$$a_i = \frac{1}{n} - b_i \bar{x} \tag{13}$$

2- Quelques identités matricielles
On considère le modèle de régression multivarié où $Y \in \mathbb{R}^n$ est la v.a. "expliquée", $X \in \mathbb{M}_{n,p+1}(\mathbb{R})$ est la matrice étendue des variables "explicative". On note $\beta \in \mathbb{R}^{p+1}$ et $\varepsilon \in \mathbb{R}^n$ l'erreur, tel que

$$Y = X\beta + \varepsilon. \tag{14}$$

1) Rappeler l'expression matricielle de l'estimateur $\hat{\beta}$ en fonction des données.
2) Montrer que $e \perp X$, $e = Y - X\hat{\beta}$.
3) Montrer que $\hat{\beta}$ réalise le minimum de la fonction

$$\min_{\gamma \in \mathbb{R}^{p+1}} \|Y - X\gamma\| \tag{15}$$

4) Montrer que $E(\hat{\beta}|X) = \beta$.
5) On note H la matrice

$$H = X(X^T X)^{-1} X^T. \tag{16}$$

Vérifier que
(i) $e = (I - H)Y$.
(ii) H est symétrique.
(iii) H est idempotente, i.e., $H^2 = H$, de même que $I - H$.
(iv) $HX = X$.
(v) $e = Y - HY$ est orthogonal à X.
(vi) $(I - H)X = 0$.
(vii) $(I - H)H = H(I - H) = 0$.

3- Effet secondaire d'un médicament
Un médicament antidépresseur est suspecté de diminuer la vigilance. On conduit une étude sur un échantillon de 10 patients qui prennent ce médicament. La vigilance est mesurée par le temps mis par le patient pour presser un bouton dès qu'il reçoit un signal.

Patient i	1	2	3	4	5	6	7	8	9	10
Dose x_i	1	5	3	8	2	2	10	8	7	4
Temps y_i	1	6	1	6	3	2	8	5	6	2

1) Tracer un diagramme x/y. Que peut-on déduire de ce diagramme ?
2) Evaluer la droite de régression moindres carrés correspondant à ces données. On a ici

$$\sum_{i=1}^{10} x_i = 50, \quad \sum_{i=1}^{10} x_i^2 = 336 \tag{17}$$

$$\sum_{i=1}^{10} y_i = 40, \quad \sum_{i=1}^{10} y_i^2 = 216 \tag{18}$$

$$\sum_{i=1}^{10} x_i y_i = 262 \tag{19}$$

3) Quel temps de réaction pronostiquez-vous pour un patient qui prend une dose de $5.5mg$?
4) Précisez un test statistique permettant de diagnostiquer si le dosage du médicament a un effet sur le temps de réaction. Formuler la question en terme de test statistique. Effectuer le test dans le cas présent au risque 5%. Interpréter les résulats.
5) Donner un intervalle de confiance pour la vigilance obtenue pour une dose de $y_0 = 5.5mg$.

4- Modèle de moindres carrés en économie
L'hypothèse du revenue permanent de M. Friedman (1957) fait dépendre la consommation au temps t du revenu et de la consommation durant la période précédente, c'est-à-dire:

$$C_t = \beta_1 Y_t + \beta_2 C_{t-1} + \epsilon_t \tag{20}$$

Les estimations des paramètres de ce modèle de régression multiple sont données dans le tableau suivant:

	$\hat{\beta}_j$	$\hat{\sigma}_{\hat{\beta}_j}$
Revenu	0.0700088	0.0144448
Consommation	0.9239275	0.0159818

Ces données sont déduites des 102 données trimestrielles du premier trimestre de 1969 au deuxième trimestre de 1990
1) Donner un intervalle de confiance de β_1 et β_2 avec un niveau de confiance de 95%. Les deux covariables sont-elles significatives au niveau $\alpha = 0.05$?
Qu'en déduit-on à propos du modèle de Friedmann ?
2) Quelle prédiction de la consommation faites-vous pour le troisième trimestre de 1990 ? ($t = 103$) pour un revenu de 6.4 et une consommation à $t = 102$ de 5.7 ?
3) Quelle(s) hypothèses habituelles d'un modèle moindres carrés est (sont) invalide(s) ?

5- Mesures de pollution
On mesure dans une grande ville sur 14 jours consécutifs la concentration en dioxyde de soufre Y ainsi que la température moyenne de la journée X_1. De plus on note si le jour est un samedi ou un dimanche, afin de déterminer si il y a un effet de week-end (lié par exemple a des départs en voiture). On note $X_2 = 1$ si le jour est un samedi ou un dimanche.

y	-3.15	-2.83	-3.02	-3.08	-3.54	-2.98	-2.78	-3.35	-2.76	-1.90	-2.12	-2.45	-1.97	-2.23
x_1	16.47	16.02	16.81	22.87	21.68	21.23	20.55	18.32	15.96	15.36	12.47	12.46	11.77	11.72
x_2	0	0	0	1	1	0	0	0	0	0	1	1	0	0

On a

$$(X^TX)^{-1} = \begin{pmatrix} 1.5488742 & -0.0882330 & -0.0162669 \\ -0.0882330 & 0.0053732 & -0.0050992 \\ -0.0162669 & 0.0050992 & 0.3548391 \end{pmatrix} \tag{21}$$

and

$$X^Ty = \begin{pmatrix} -38.16486 \\ -656.46618 \\ -11.19324 \end{pmatrix} \tag{22}$$

1) Quelle est l'estimation des coefficients de la régression linéaire multiple ? Interprétation ?
2) Le coefficient de détermination est

$$R = 0.5781 \tag{23}$$

Est-ce que les facteurs explicatifs sont pertinents pour expliquer la concentration en dioxyde de soufre ? Effectuer un test au niveau $\alpha = 0.01$.
3) On obtient comme estimations des variances $\sigma_{\hat{\beta}_1}$ et $\sigma_{\hat{\beta}_2}$ les valeurs

$$\hat{\sigma}_{\hat{\beta}_1} = 0.0267, \quad \hat{\sigma}_{\hat{\beta}_2} = 0.2169 \tag{24}$$

Tester l'hypothèse $\beta_i = 0$, $i = 1, 2$ au niveau $\alpha = 0.05$. Effectuer une régression linéaire à l'aide de celle des deux covariables qui a la plus grande influence.

6- Etudes de cerisiers
Le fichier `cherry.txt` contient la description de 3 caractéristiques de 32 cerisiers. On souhaite effectuer en matlab une analyse statistique multivariée de type régression de ces données.
1) Lire les données dans des tableaux `volume`, `diametre`, `hauteur`.
2) On s'intéresse à l'hypothèse suivante: existe-t-il un modèle statistique de type régresssion linéaire de la forme

$$y_i = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + e_i \tag{25}$$

avec

- y_i: le volume de l'arbre i.
- $x_{i,1}$: le diamètre de l'arbre i.
- $x_{i,2}$: la hauteur de l'arbre i.
- e_i : variables aléatoires indépendantes de loi $N(0, \sigma^2)$?

Représenter sur la Fig.1 le scatterplot des données sous la forme volume (y)/ diamètre(x). Représenter sur la Fig.2 volume(y)/hauteur(x). Est-ce que ces figures permettent de valider une hypothèse de régression du type (25) ?
3) Evaluer les estimations des 4 paramètres $\hat{\beta}_0$, $\hat{\beta}_1$, $\hat{\beta}_2$ à l'aide de la fonction `regress`. (On obtient $\hat{\beta}_0 = -44.87706$, $\hat{\beta}_1 = 5.1606$, $\hat{\beta}_2 = 0.0945$).
4) Utiliser la routine `regstats` pour évaluer les informations suivantes:

- estimation de la variance sur les valeurs $\hat{\beta}_0$, $\hat{\beta}_1$, $\hat{\beta}_2$.
- estimation de la variance $\hat{\sigma}_e^2$ du modèle moindres carrés.
- calcul des niveaux de confiance de type "p" conduisant à un rejet de l'hypothèse $(H_0) : \beta_i = 0$, $i = 0, 1, 2$ au vu des données.

5) Tracer sur la Fig.3 les résidus en fonction du diamètre, puis sur la Fig.4 , les résidus en fonction de la hauteur.
6) Tracer sur la Fig.7 le normalplot des résidus. On utilisera la fonction `qqplot`.
7) Si on assimile un arbre à un cône de diamètre d et de de hauteur h, alors son volume est donné explictement par

$$v = \frac{\pi}{12} h d^2. \tag{26}$$

Vérifier que la formule (26) justifie de s'intéresser à un modèle logarithmique de régression linéaire de la forme

$$\ln(v) = \beta_0 + \beta_1 \ln(h) + \beta_2 \ln(d). \tag{27}$$

8) Reprendre le programme de travail des questions précédentes: on commencera par tracer sur la Fig.8 le logarithme du volume(y)/logarithme du diamètre(x), puis, sur la Fig.9, logarithme du volume (y)/logarithme de la hauteur (x). Evaluer ensuite $\hat{\beta}_0$, $\hat{\beta}_1$, $\hat{\beta}_2$ et σ^2 ainsi que les diagnostics statistiques associés.
9) Vérifier que l'arbre numéro 26 possède des caractéristiques de type "aberrantes". On tracera des box-plots sur la Fig.10. On tracera également le graphe de la distance de Cook.
10) Reprendre les calculs précédents sans l'arbre 26 (en version logarithmique).
11) Effectuer une estimation de l'augmentation de volume d'un arbre lorsque le diamètre augmente de 25 %. On donnera un intervalle de confiance pour cette estimation.