Université Sultan Moulay Slimane
Faculté des Sciences et Techniques
Béni Mellal
Département Informatique

TP N°2----REGRESSION LINEAIRE

La fonction de la régression linéaire sous R est lm (lineaire model) :

- lm(x~y) : y = ax+b

Exercice 1 :

- **Créer les variables quantitatives x et y**

```
>  xi <- c(1:12)
> yi <- c(40, 42, 44, 45, 48, 50, 52, 55, 58, 63, 68, 70)
```

- **Tracer le nuage de points (x,y)**

```
>  plot(xi, yi)
```

- **coefficient de correlation**

```
r<-cor(xi,yi)
cor.test(xi,yi)
## autrement
mat<- data.frame(xi,yi)
#source("http://www.sthda.com/upload/rquery_cormat.r")
require("corrplot")
rquery.cormat(mat)
```

- **condition normalité**

```
  #shapiro-wilk ou kolmogorov smirnov
ks.test(yi,"pnorm")
ks.test(xi,"pnorm")
shapiro.test(xi)
shapiro.test(yi)
```

- **Etablir la régression linéaire entre x et y**

```
> regxy <- lm(yi ~ xi)
```

on obtient :

```
Call:
lm(formula = yi ~ xi)
```

Coefficients:

(Intercept) xi

35.076 2.745

Donc l'équation de régression est : y = 2,74x + 35,08

- **Pour ajouter la droite de régression sur le nuage de points**

  > abline(35.076, 2.745, col = 'red')

- **Pour plus de détails sur le modèle**

  > summary(lm(yi ~ xi))

Call:

lm(formula = yi ~ xi)

Residuals:

Min 1Q Median 3Q Max

-2.2890 -1.6029 -0.1614 1.5728 2.7319

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 35.0758 1.1612 30.2 3.70e-11 ***

xi 2.7448 0.1578 17.4 8.35e-09 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.887 on 10 degrees of freedom

Multiple R-squared: 0.968, Adjusted R-squared: 0.9648

F-statistic: 302.6 on 1 and 10 DF, p-value: 8.355e-09

- **Tester l'indépendance des résidus**

  acf(residuals(regxy), main="reg xy")

  #

  > library(car)

  > durbinWatsonTest (regxy) ou

  # dwtest(regxy)

- **Tester la normalité**

  plot(regxy,2)

  shapiro.test(residuals(regxy))

  residualPlot((regxy))

  residualPlots((regxy))

- **Tester l'homogénéité des résidus**

  plot(regxy, 3)

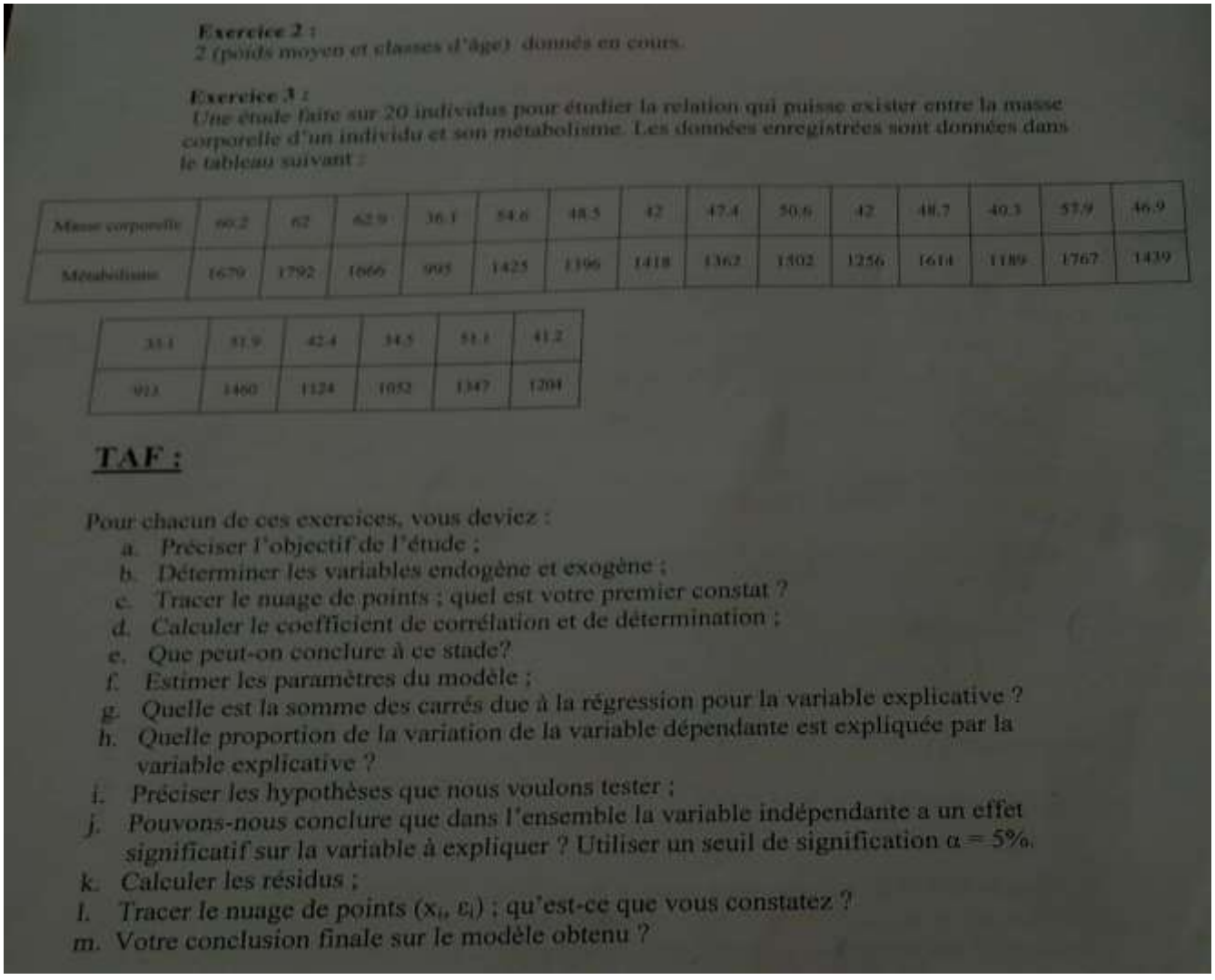
ncvTest(regxy)

- **Intervalle de confiance**

  plot(regxy,1)

  confint(regxy)

Interpréter les résultats de chaque exécution.



Exercice 2 :
2 (poids moyen et classes d'âge) : donnés en cours.

Exercice 3 :
Une étude faite sur 20 individus pour étudier la relation qui puisse exister entre la masse corporelle d'un individu et son métabolisme. Les données enregistrées sont données dans le tableau suivant :

Masse corporelle	60.2	62	62.9	36.1	54.6	48.5	42	47.4	50.6	42	48.7	40.3	57.9	46.9
Métabolisme	1679	1792	1666	995	1425	1396	1418	1362	1502	1256	1614	1189	1767	1439

33.1	51.9	42.4	34.5	51.1	41.2
913	1460	1124	1052	1347	1204

TAF :

Pour chacun de ces exercices, vous deviez :

a. *Préciser l'objectif de l'étude ;*
b. *Déterminer les variables endogène et exogène ;*
c. *Tracer le nuage de points ; quel est votre premier constat ?*
d. *Calculer le coefficient de corrélation et de détermination ;*
e. *Que peut-on conclure à ce stade?*
f. *Estimer les paramètres du modèle ;*
g. *Quelle est la somme des carrés due à la régression pour la variable explicative ?*
h. *Quelle proportion de la variation de la variable dépendante est expliquée par la variable explicative ?*
i. *Préciser les hypothèses que nous voulons tester ;*
j. *Pouvons-nous conclure que dans l'ensemble la variable indépendante a un effet significatif sur la variable à expliquer ? Utiliser un seuil de signification $\alpha = 5\%$.*
k. *Calculer les résidus ;*
l. *Tracer le nuage de points (x_i, ε_i) ; qu'est-ce que vous constatez ?*
m. *Votre conclusion finale sur le modèle obtenu ?*

II. Régression multiple :

Exercice 1

Cet exercice est issu du livre « Probabilités, Statistique et technique de régression » de Gérald Baillargeon. Les éditions SMG

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. Nous voulons examiner si les facteurs suivants peuvent influencer 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 due 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. Pouvons-nous conclure que dans l'ensemble les trois variables explicatives ont un effet significatif sur le niveau d'anxiété ? Utiliser un seuil de signification $\alpha = 5\%$.
Préciser les hypothèses que nous voulons tester.
4. Si nous ne tenons compte que de la variable explicative X_1, quel serait alors le tableau d'analyse de la variance correspondant ?
5. Donner les différents tests d'hypothèses qu'on peut effectuer ;
6. Tester ces hypothèses au seuil de signification $\alpha = 5\%$, en utilisant un rapport F approprié :
7. Quelle est la valeur du coefficient de détermination R2 associée à l'estimation de chaque modèle spécifié à la question 5. ?
8. Lequel des trois modèles semble le mieux approprié pour expliquer les fluctuations du niveau d'anxiété des cadres d'entreprises ?

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_i	Épaisseur du matériau X_{i1}	Densité X_{i2}
1	37,8	4	4,0
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

Même questions que l'exercice 1.

Exercice 3

Les données à traiter pour la RLM sont fournies dans le tableau ci-dessous. Il s'agit de l'exemple repris de la chenille processionnaire du pin traité dans l'ouvrage de TOMASSONE & al. Le fichier de données est composé de 33 placettes où sont plantés des arbres infectés par des nids de chenille « processionnaire du pin », une variable réponse (X11 et sa transformée en Log et dix variables régresseurs potentiels (X1-X10).

« Les expérimentateurs souhaitent connaître l'influence de certaines caractéristiques e peuplements forestiers (variables régresseurs X1-X10) sur le développement de la chenille processionnaire du pin (variable réponse X11 ou son logarithme ».

Données Chenille processionnaire de TOMASSONE

Obs	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10	X11	log
1	1200	22	1	4.0	14.8	1.0	1.1	5.9	1.4	1.4	2.37	0.86289
2	1342	28	8	4.4	18.0	1.5	1.5	6.4	1.7	1.7	1.47	0.38526
3	1231	28	5	2.4	7.8	1.3	1.6	4.3	1.5	1.4	1.13	0.12222
4	1254	28	18	3.0	9.2	2.3	1.7	6.9	2.3	1.6	0.85	-0.16252
5	1357	32	7	3.7	10.7	1.4	1.7	6.6	1.8	1.3	0.24	-1.42712
6	1250	27	1	4.4	14.8	1.0	1.7	5.8	1.3	1.4	1.49	0.39878
7	1422	37	22	3.0	8.1	2.7	1.9	8.3	2.5	2.0	0.30	-1.20397
8	1309	46	7	5.7	19.6	1.5	1.3	7.8	1.8	1.6	0.07	-2.65926
9	1127	24	2	3.5	12.6	1.0	1.7	4.9	1.5	2.0	3.00	1.09861
10	1075	34	9	4.3	12.0	1.6	1.8	6.8	2.0	2.0	1.21	0.19062
11	1166	24	17	5.5	16.7	2.4	1.5	11.5	2.9	1.7	0.38	-0.96758
12	1182	41	32	5.4	21.6	3.3	1.4	11.3	2.8	2.0	0.70	-0.35667
13	1179	15	0	3.2	10.5	1.0	1.7	4.0	1.1	1.6	2.64	0.97078
14	1256	21	0	5.1	19.5	1.0	1.8	5.8	1.1	1.4	2.05	0.71784
15	1251	26	2	4.2	16.4	1.1	1.7	6.2	1.3	1.8	1.75	0.55962
16	1536	38	31	5.7	17.8	3.1	1.7	11.4	2.8	1.9	0.06	-2.81341
17	1554	27	20	5.6	20.2	2.8	1.9	9.2	2.7	1.3	0.13	-2.04022
18	1305	30	6	3.8	15.7	1.4	1.2	7.2	2.1	1.9	1.00	0.00000
19	1316	34	8	3.1	11.4	1.5	1.8	5.0	1.6	2.0	0.41	-0.89160
20	1427	39	19	4.6	15.2	2.4	1.6	9.1	2.4	1.9	0.72	-0.32850
21	1575	20	32	5.2	18.9	3.0	1.7	9.4	2.5	1.8	0.67	-0.40048
22	1397	26	16	4.2	14.8	2.2	1.6	7.7	2.2	1.8	0.12	-2.12026
23	1377	29	4	5.3	19.8	1.2	1.8	6.8	1.6	1.9	0.97	-0.03046
24	1574	24	23	5.2	17.8	2.4	1.8	7.8	2.2	2.0	0.07	-2.65926
25	1396	45	13	4.7	15.2	1.7	1.6	7.8	2.1	1.4	0.10	-2.30259
26	1393	27	5	4.7	18.3	1.2	1.7	7.5	1.7	2.0	0.68	-0.38566
27	1433	23	18	6.5	21.0	2.7	1.8	13.7	2.7	1.3	0.13	-2.04022
28	1349	24	1	2.7	5.8	1.0	1.7	3.6	1.3	1.8	0.20	-1.60944
29	1208	23	2	3.5	11.5	1.1	1.7	5.4	1.3	2.0	1.09	0.08618
30	1198	28	15	3.9	11.3	2.0	1.6	7.4	2.8	2.0	0.18	-1.71480
31	1228	31	6	5.4	21.8	1.3	1.7	7.0	1.5	1.9	0.35	-1.04982
32	1229	21	11	5.8	16.7	1.7	1.8	10.0	2.3	2.0	0.21	-1.56065
33	1310	36	17	5.2	17.8	2.3	1.9	10.3	2.6	2.0	0.03	-3.50656

3

Avec :

X11 : Nombre de nids de processionnaires par arbre d'une placette.
Log = Log(X11), transformation de la variable X11 par son logarithme

X1 : Altitude (en mètre)
X2 : pente (en degré)
X3 : nombre de pins dans une placette de 5 ares
X4 : hauteur de l'arbre échantillonné au centre de la placette
X5 : diamètre de cet arbre

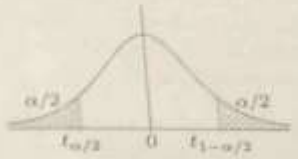
1. Mêmes questions (a-d) de l'exercice 1 (cas simple) ;
2. Qu'est ce que vous constatez ?

A.3. Lois de Student

Si T est une variable aléatoire suivant la loi de Student à ν degrés de liberté, la table donne, pour α fixé, la valeur $t_{1-\alpha/2}$ telle que

$$P[|T| \geqslant t_{1-\alpha/2}] = \alpha.$$

Ainsi, $t_{1-\alpha/2}$ est le quantile d'ordre $1-\alpha/2$ de la loi de Student à ν degrés de liberté.



ν \ α	0,900	0,500	0,300	0,200	0,100	0,050	0,020	0,010	0,001
1	0,1584	1,0000	1,9626	3,0777	6,3138	12,7062	31,8205	63,6567	636,6193
2	0,1421	0,8165	1,3862	1,8856	2,9200	4,3027	6,9646	9,9248	31,5991
3	0,1366	0,7649	1,2498	1,6377	2,3534	3,1824	4,5407	5,8409	12,9240
4	0,1338	0,7407	1,1896	1,5332	2,1318	2,7764	3,7469	4,6041	8,6103
5	0,1322	0,7267	1,1558	1,4759	2,0150	2,5706	3,3649	4,0321	6,8688
6	0,1311	0,7176	1,1342	1,4398	1,9432	2,4469	3,1427	3,7074	5,9588
7	0,1303	0,7111	1,1192	1,4149	1,8946	2,3646	2,9980	3,4995	5,4079
8	0,1297	0,7064	1,1081	1,3968	1,8595	2,3060	2,8965	3,3554	5,0413
9	0,1293	0,7027	1,0997	1,3830	1,8331	2,2622	2,8214	3,2498	4,7809
10	0,1289	0,6998	1,0931	1,3722	1,8125	2,2281	2,7638	3,1693	4,5869
11	0,1286	0,6974	1,0877	1,3634	1,7959	2,2010	2,7181	3,1058	4,4370
12	0,1283	0,6955	1,0832	1,3562	1,7823	2,1788	2,6810	3,0545	4,3178
13	0,1281	0,6938	1,0795	1,3502	1,7709	2,1604	2,6503	3,0123	4,2208
14	0,1280	0,6924	1,0763	1,3450	1,7613	2,1448	2,6245	2,9768	4,1405
15	0,1278	0,6912	1,0735	1,3406	1,7531	2,1314	2,6025	2,9467	4,0728
16	0,1277	0,6901	1,0711	1,3368	1,7459	2,1199	2,5835	2,9208	4,0150
17	0,1276	0,6892	1,0690	1,3334	1,7396	2,1098	2,5669	2,8982	3,9651
18	0,1274	0,6884	1,0672	1,3304	1,7341	2,1009	2,5524	2,8784	3,9216
19	0,1274	0,6876	1,0655	1,3277	1,7291	2,0930	2,5395	2,8609	3,8834
20	0,1273	0,6870	1,0640	1,3253	1,7247	2,0860	2,5280	2,8453	3,8495
21	0,1272	0,6864	1,0627	1,3232	1,7207	2,0796	2,5176	2,8314	3,8193
22	0,1271	0,6858	1,0614	1,3212	1,7171	2,0739	2,5083	2,8188	3,7921
23	0,1271	0,6853	1,0603	1,3195	1,7139	2,0687	2,4999	2,8073	3,7676
24	0,1270	0,6848	1,0593	1,3178	1,7109	2,0639	2,4922	2,7969	3,7454
25	0,1269	0,6844	1,0584	1,3163	1,7081	2,0595	2,4851	2,7874	3,7251
26	0,1269	0,6840	1,0575	1,3150	1,7056	2,0555	2,4786	2,7787	3,7066
27	0,1268	0,6837	1,0567	1,3137	1,7033	2,0518	2,4727	2,7707	3,6896
28	0,1268	0,6834	1,0560	1,3125	1,7011	2,0484	2,4671	2,7633	3,6739
29	0,1268	0,6830	1,0553	1,3114	1,6991	2,0452	2,4620	2,7564	3,6594
30	0,1267	0,6828	1,0547	1,3104	1,6973	2,0423	2,4573	2,7500	3,6460
40	0,1265	0,6807	1,0500	1,3031	1,6839	2,0211	2,4233	2,7045	3,5510
60	0,1262	0,6786	1,0455	1,2958	1,6706	2,0003	2,3901	2,6603	3,4602
80	0,1261	0,6776	1,0432	1,2922	1,6641	1,9901	2,3739	2,6387	3,4163
120	0,1259	0,6765	1,0409	1,2886	1,6577	1,9799	2,3578	2,6174	3,3735
∞	0,1257	0,6745	1,0364	1,2816	1,6449	1,9600	2,3263	2,5758	3,2905