Licence de Psychologie - UV PSY38X2
Corrigé de l'épreuve de Traitement des données en Psychologie donnée en juin 2003

Exercice 1 – Informatique (6 points)

L'association étudiante dont vous faites partie collabore avec une crèche-garderie pour accueillir les enfants des étudiant(e)s. L'association est notamment chargée de produire des plannings donnant, pour chaque demi-journée, la liste des enfants qui seront confiés à la crèche.
L'accueil se fait par demi-journée (lundi matin, lundi après-midi, etc). L'enfant est inscrit pour certaines demi-journées, pendant une période définie par une date de début et une date de fin (par exemple, tous les lundis matin, pendant la période du 17 février 2003 au 19 mai 2003).
Pour chaque enfant, on veut connaître son nom et son prénom, ainsi que le nom et le prénom de l'un des parents. On veut également disposer de la liste des demi-journées pour lesquelles il a été inscrit à la crèche.
1) Faire une liste des données qui apparaissent dans l'énoncé ci-dessus.
L'énoncé fait apparaître les données suivantes : demi-journée, début-période, fin-période, nom-enfant, prénom-enfant, nom-parent, prénom-parent.
2) On décide de répartir ces données entre trois entités : *Parent*, *Enfant*, *Inscription* et deux associations binaires "*... est responsable de...*" et "*... est accueilli pour la demi-journée du ...*". Répartir les données précédentes comme attributs de ces entités et associations. Indiquer (ou proposer) un identifiant pour chacune des entités.
Pour chacune des entités "Parent" et "Enfant", on va ajouter un attribut (respectivement Code-P et Code-E) qui jouera le rôle d'identifiant. Ces entités seront donc définies par :

Parent
Code-P
nom-parent
prénom-parent

Enfant
Code-E
nom-enfant
prénom-enfant

La donnée "demi-journée" se rattache à l'entité "Inscription". Il peut être commode (mais pas obligatoire) d'ajouter un code tel que "Lu1" pour lundi matin, "Lu2" pour lundi après-midi, etc par exemple. Ainsi, l'entité "Inscription" est définie par :

Inscription
Code-I
demi-journée

L'association "*... est responsable de...*" relie les deux entités "Parent" et "Enfant" et ne possède pas d'attribut. L'association "*... est accueilli pour la demi-journée du ...*" relie les entités "Enfant" et "Inscription" et possède comme attributs les données début-période et fin-période.
3) Représenter à l'aide d'un schéma le modèle conceptuel des données précédent.
Cf. figure 1
4) L'association a déjà constitué une table Parents-Enfants, de structure :

Parents-Enfants(Nom, Prénom, Nom-enfant, Prénom-enfant, Date-Naissance).

On a représenté ci-dessous un extrait de cette table :

Nom	Prénom	Nom-enfant	Prénom-Enfant	Date-Nais.
MARTIN	Marie	MARTIN	Pierre	01/04/01
DUPOND	Dominique	DUPOND	Sophie	15/01/00
DUPONT	Jacques	DUVAL	Laure	23/02/00
DUPOND	Dominique	PETIT	Aurélie	15/04/01
DUPOND	Dominique	DUPOND	Pierre	24/03/01

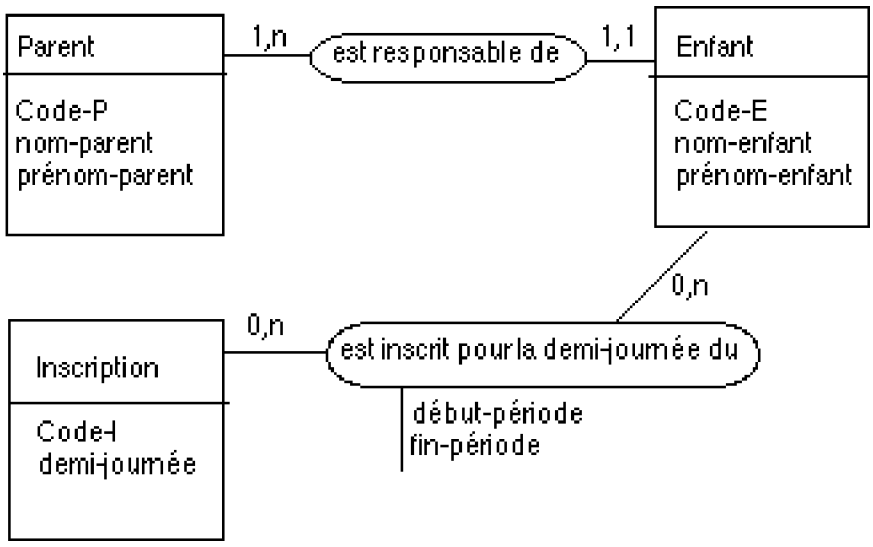

Figure 1: Modèle conceptuel des données

a) Ainsi structurée, une telle table est ambiguë. Pourquoi? Quel attribut proposeriez-vous d'ajouter pour résoudre ce problème?
Au vu de cette table, il est impossible de déterminer si "DUPOND Dominique", par exemple, désigne une ou plusieurs personnes. Cette table est donc ambiguë. Pour résoudre ce problème, il faudrait introduire un "code-parent" identifiant sans ambiguïté un parent donné.
b) En rassemblant ainsi l'ensemble des données dans une seule table, est-il possible d'éviter les redondances (données figurant plusieurs fois dans la base)?
En rassemblant les données en une seule table, il est impossible de ne pas répéter les coordonnées d'un parent plusieurs fois s'il a plusieurs enfants inscrits.
Proposer une structure en deux tables permettant de mieux représenter les données.
Il vaudrait mieux répartir les données en une table "Parents", d'attributs Code-parent, Nom, Prénom, et une table "Enfants", avec comme attributs: Code-enfant, Nom-enfant, Prénom-enfant, Date-naissance, Code-parent. Ce dernier attribut représente l'information "tel parent est responsable de tel enfant" : c'est une clé étrangère pour la table "Enfants".
Quelle opération de l'algèbre relationnelle permettra alors d'obtenir la table précédente?
La table précédente pourra ensuite être obtenue en effectuant une jointure des tables "Parents" et "Enfants" sur l'attribut commun "Code-parent".

Exercice 2 – Statistiques (9 points)

Dans un article publié en 2003, trois chercheurs du Laboratoire de Psychologie Expérimentale de l'Université René Descartes rendent compte d'une expérimentation qu'ils ont menée concernant l'utilisation de produits multimédias pour l'apprentissage.
La multiplicité des types d'informations et la multimodalité semblent avoir assis la popularité des produits multimédia sans toutefois garantir leur efficacité.
La charge cognitive qu'impose l'apprentissage multimédia varierait principalement selon la quantité d'informations et le mode sensoriel dans lequel elles sont présentées. Dans ce cadre, présenter des textes dans le mode auditif et des diagrammes dans le mode visuel est supposé conditionner leur prise en charge par deux sous-systèmes distincts de la mémoire de travail. Au contraire, la présentation de ces différentes informations dans le même mode aurait pour conséquence la surcharge de l'un des sous-systèmes de la mémoire de travail. L'objectif de la présente expérience est d'étudier la mise en jeu des composantes visuo-spatiale et verbale de la mémoire de travail

dans la compréhension et le maintien en mémoire d'informations multimédia.
Cinquante-six élèves ingénieurs ont participé à l'expérience. Six séquences multimédias ont été développées dans deux formats différents : dans un format, les textes et les indications sont présentés auditivement (AA), dans l'autre, les textes sont présentés visuellement et les indications auditivement (VA). Dans chacun des cas, les séquences comprennent en outre la présentation visuelle de diagrammes.
Vingt-huit sujets ont vu les séquences dans le format AA et vingt-huit dans le format VA. Dans chaque cas, deux séquences étaient visionnées sans tâche concurrente (condition contrôle), deux autres avec une tâche concurrente spatiale et les deux dernières avec une tâche concurrente verbale.
Afin d'évaluer le maintien en mémoire des informations contenues dans les séquences multimédias, une épreuve de reconnaissance a été mise au point. Elle se compose pour moitié d'items mettant en jeu la reconnaissance d'un diagramme, et pour moitié d'items mettant en jeu la reconnaissance d'une phrase.
Pour chaque groupe de deux séquences et chaque type de reconnaissance (diagramme v/s phrase), le score du sujet est le pourcentage de réponses correctes.

1) Etude du plan d'expérience utilisé.
a) Quels sont les facteurs de variation qui sont pris en compte ? Quels sont les niveaux de chacun d'eux ?
Les facteurs pris en compte dans l'expérience sont :

- Le facteur "format de présentation", à deux niveaux AA et VA : $\mathcal{F}_2$;
- Le facteur "condition de passation", à trois niveaux (condition contrôle, tâche concurrente spatiale, tâche concurrente verbale) : $\mathcal{C}_3$;
- Le facteur "type de reconnaissance", à deux niveaux : $\mathcal{R}_2$;
- Le facteur "sujet", à 56 niveaux : $\mathcal{S}_{56}$.

b) Ecrire les relations d'emboîtement ou de croisement entre le facteur "sujet" et chacun des autres facteurs, puis entre les autres facteurs.
Les sujets ont été répartis en deux groupes selon le format de présentation. Le facteur "sujet" est donc emboîté dans le facteur $\mathcal{F}_2$: $\mathcal{S}_{28} < \mathcal{F}_2 >$.
Chaque sujet est soumis aux trois conditions de passation et aux deux types de de reconnaissance. Chacun de ces facteurs est croisé avec le facteur sujet : $\mathcal{S}_{56} * \mathcal{C}_3$; $\mathcal{S}_{56} * \mathcal{R}_2$.
Les facteurs "format de présentation", "condition de passation" et "type de reconnaissance" sont croisés deux à deux : $\mathcal{F}_2 * \mathcal{C}_3$; $\mathcal{F}_2 * \mathcal{R}_2$; $\mathcal{C}_3 * \mathcal{R}_2$.
c) Quel est le plan d'expérience utilisé ?
L'expérience a été menée selon le plan : $\mathcal{S}_{28} < \mathcal{F}_2 > * \mathcal{C}_3 * \mathcal{R}_2$.

2) Sur les figures 2 et 3, on a indiqué les moyennes des scores des sujets selon les conditions. Utiliser les données fournies par ces figures pour construire un diagramme d'interaction entre les facteurs "tâche" et "format", dans la condition "reconnaissance de phrases".
Cf. figure 4.
3) On étudie dans un premier temps les résultats observés dans la condition contrôle. On soumet les données correspondantes à une analyse de variance.
Le tableau d'analyse de variance se présente ainsi :

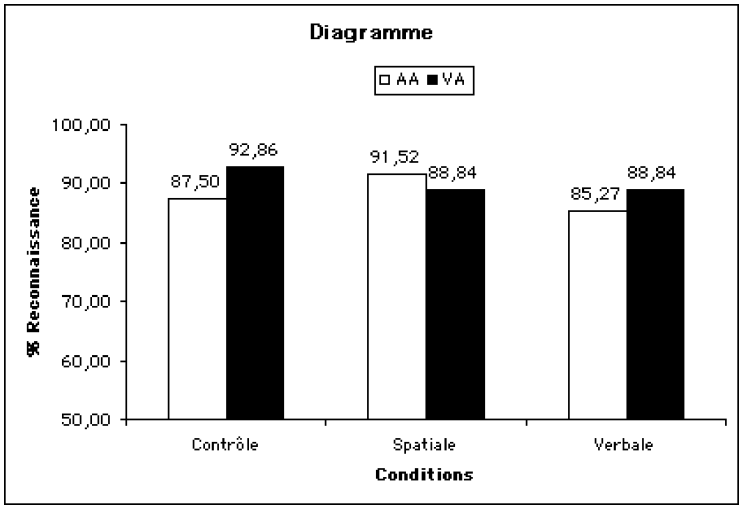

Figure 2: Reconnaissance de diagrammes

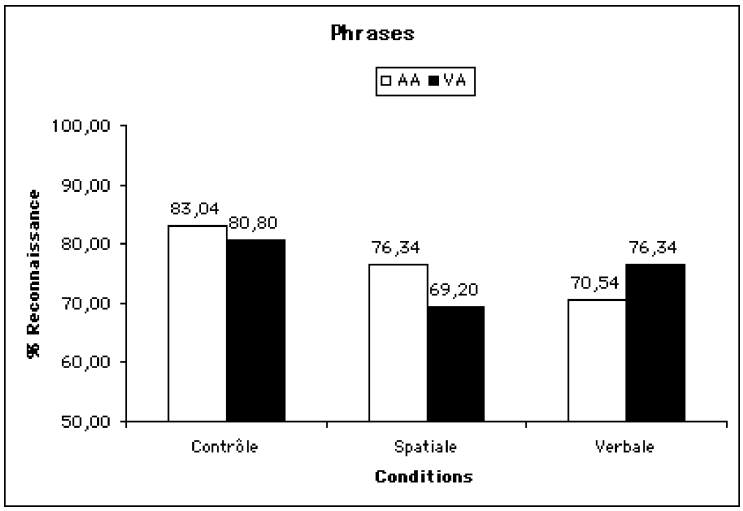

Figure 3: Reconnaissance de phrases

Sources de var.	ddl	SC	CM	F
Format	1	68.14	...	...
Résidu "inter-sujets"	54	4320.10	...	
Reconnaissance	1	1910.37	...	...
Rec.×Format	1	404.32	...	...
Résidu "intra-sujets"	54	10496.77		
Total	111	17199.70		

a) Quel est le plan d'expérience correspondant à cette partie de l'étude?
On se limite ici à la modalité "condition contrôle" du facteur $\mathcal{C}_2$. Le plan correspondant est donc : $\mathcal{S}_{28} < \mathcal{F}_2 > *\mathcal{R}_2$.
b) Compléter ce tableau en calculant les degrés de liberté, les carrés moyens et les statistiques F de Fisher qui sont remplacés par "..." dans le tableau ci-dessus.

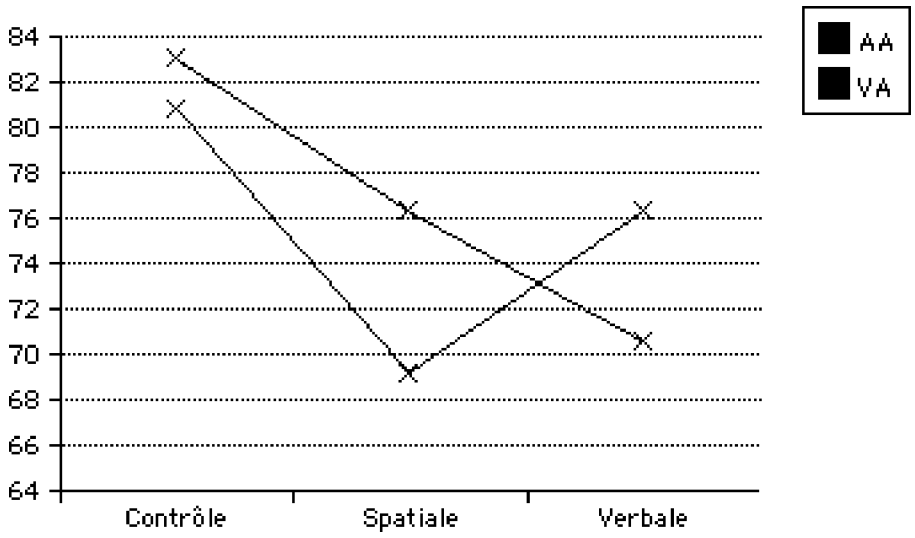

Figure 4: Diagramme d'interaction

Sources de var.	ddl	SC	CM	F
Format	1	68.14	68.14	0.85
Résidu "inter-sujets"	54	4320.10	80.00	
Reconnaissance	1	1910.37	1910.37	9.83
Rec.×Format	1	404.32	404.32	2.08
Résidu "intra-sujets"	54	10496.77	194.38	
Total	111	17199.70		

c) En utilisant un seuil de 5%, étudier quelles sont les sources de variation dont l'effet est significatif.
Pour chacune des trois sources de variation, les nombres de degrés de liberté à prendre en compte sont : $ddl_1 = 1$ et $ddl_2 = 54$. Au seuil de 5%, on a : $F_{crit}(1, 54) = 4.02$.
Pour le facteur $\mathcal{F}_2$, on a $F_{calc} = 0.85$ et donc $F_{calc} < F_{crit}$. Le facteur "format de présentation" n'a donc pas d'effet significatif.
De même, l'effet de l'interaction $\mathcal{R}_2 \times \mathcal{F}_2$ n'est pas significatif.
En revanche, pour le facteur $\mathcal{R}_2$, on a $F_{calc} = 9.83$ et donc $F_{calc} > F_{crit}$. L'effet du facteur "Type de reconnaissance" est donc significatif au seuil de 5%.

4) Pour la condition "tâche concurrente spatiale", les auteurs indiquent :

> Considérons à présent les effets de la tâche concurrente spatiale. L'analyse statistique révèle que sa présence fait globalement chuter les performances de 86.05% (dans la condition contrôle) à 81.47% [1]. L'effet est significatif ($F(1, 54) = 5.59$; $p < .025$). L'examen statistique révèle que cet effet délétère ne varie pas selon le format multimédia ($F(1, 54) = 2.79$)[2]. Toutefois l'examen de comparaisons spécifiques indique que la tache de "tapping" fait significativement chuter les performances dans le format VA (de 86.83 % dans 1a condition controle à 79.02% ($F(1, 27) = 11.23$; $p < .0025$) mais pas dans le format AA ($F \leq 1$) ce qui est conforme à nos hypothèses [3]. D'autre part, l'analyse statistique indique que l'effet néfaste de la tâche concurrente spatiale varie selon les items de l'épreuve de reconnaissance ($F(1, 54) = 5.19$; $p < .05$) [4]. Contrairement à ce que l'on aurait pu attendre la tâche concurrente spatiale n'affecte pas les performances de reconnaissance de diagrammes ($F < 1$) [5]. Par contre, elle altère significativement la reconnaissance de phrases comme on peut le

constater sur la figure 3 ($F(1,54) = 7.21\,;\, p < .025$) [6]. Précisément, elle détériore la reconnaissance de phrases uniquement dans le format VA ($F1,27) = 7.25\,;\, p < .025$), pas dans le format AA ($F(1,27) = 1.60$).

a) A l'aide des données fournies dans l'énoncé, vérifier les valeurs indiquées dans la phrase [1].
Les quatre moyennes observées dans la condition "tâche concurrente spatiale" sont : 91.52, 88.84, 76.34, 69.20 et on vérifie bien que : $\frac{91.52 + 88.84 + 76.34 + 69.20}{4} = 81.47$. Comme les quatre groupes considérés sont équilibrés (c'est-à-dire de même effectif), cette dernière valeur représente aussi la moyenne de l'ensemble des scores individuels observés dans cette condition.
De même, on vérifie que, dans la condition contrôle : $\frac{87.50 + 92.86 + 83.04 + 80.80}{4} = 86.05$.
b) Dans la phrase [3], les auteurs évoquent des "comparaisons spécifiques". A votre avis, quels sont les tests statistiques qui ont été réalisés, et sur quelles données ont-ils porté?
L'analyse a porté ici uniquement sur les résultats des 28 sujets soumis au format VA et les tests réalisés ont permis de comparer leur score en condition contrôle et leur score en condition "tâche concurrente spatiale". Les indications de l'énoncé semblent indiquer que les auteurs ont construit un protocole dérivé en calculant pour chaque sujet une moyenne des scores observés selon le type de reconnaissance. Mais ces résultats ont aussi pu être obtenus en traitant les données selon un plan : $\mathcal{S}_{28} * \mathcal{C}_2 * \mathcal{R}_2$. Ces données ont été analysées à l'aide d'une analyse de variance (utilisation de la statistique F de Fisher, et non du t de Student).
c) Définir avec précision les individus statistiques et la variable dépendante utilisée pour obtenir les résultats indiqués dans la phrase [4]? Comment les auteurs ont-ils procédé pour obtenir des résultats?
Une lecture superficielle pourrait laisser penser que les individus statistiques pris en compte ici sont les différents items de l'épreuve de reconnaissance. Mais l'indication ($F(1,54) = 5.19\,;\, p < .05$) montre que l'expression "selon les items de l'épreuve de reconnaissance" ne fait qu'opposer les deux niveaux du facteur "Type de reconnaissance" (diagramme v/s phrase). Autrement dit, les individus statistiques et la variable dépendante sont, comme précédemment, les sujets et leurs scores. Les auteurs citent "l'effet néfaste de la tâche concurrente spatiale", ce qui indique qu'ils ont comparé la condition "tâche concurrente spatiale" à la condition contrôle. Ils s'intéressent ici aux variations de l'effet de la tâche selon le type de reconnaissance. Autrement dit, ils s'intéressent ici à l'interaction $\mathcal{C}_2 \times \mathcal{R}_2$.

Exercice 3 – Statistiques (5 points)

En 1992, le niveau des dépenses de santé par personne affiliée au régime général de Sécurité Sociale variait du simple à plus du double entre les départements français. On cherche ici à construire des modèles explicatifs de ces disparités au moyen de régressions linéaires.
La variable à expliquer est l'indicateur de remboursement de soins du secteur libéral, ou IDRS (exprimé en francs).

1) Dans un premier temps, on étudie la corrélation entre la variable à expliquer et la variable "densité de médecins pour 100 000 habitants", sur un échantillon de 10 départements. Les données sont les suivantes :

Département	Densité	IDRS
Ain	123	2302
Ariège	181	3868
Cher	143	3193
Haute-Corse	223	4860
Eure	128	2774
Finistère	173	3277
Isère	192	3077
Loiret	156	3020
Marne	162	2853
Yonne	159	3497

On donne par ailleurs :
$\sum x_i = 1\,640$; $\sum y_i = 32\,721$; $\sum x_i^2 = 277\,106$; $\sum y_i^2 = 111\,466\,229$; $\sum x_i y_i = 5\,525\,739$.
a) Calculer la covariance et le coefficient de corrélation des variables "Densité" et "IDRS". La corrélation est-elle significative au seuil de 5% ?
$Cov(\text{Densité}, \text{IDRS}) = \dfrac{5525739}{10} - \dfrac{1640}{10} \times \dfrac{32721}{10} = 15949.50.$
$Var(\text{Densité}) = \dfrac{277106}{10} - \dfrac{1640^2}{10^2} = 814.60.$
$Var(\text{IDRS}) = \dfrac{111466229}{10} - \dfrac{32721^2}{10^2} = 439984.49.$
$r_{obs} = \dfrac{15949.50}{\sqrt{814.60}\sqrt{439984.49}} = 0.84.$
L'étude porte ici sur 10 individus statistiques. Le nombre de degrés de liberté à prendre en compte est donc : $ddl = 10 - 2 = 8$. Pour un seuil de 5% et un test bilatéral, on lit dans la table : $\rho_{crit} = 0.63$. Autrement dit :
Si r est compris entre -0.63 et 0.63, on retient l'hypothèse nulle $\rho = 0$, c'est-à-dire l'indépendance des deux variables sur la population.
Dans le cas contraire, on retient l'hypothèse alternative d'une corrélation non nulle entre les deux variables.
C'est ici cette seconde hypothèse qui est retenue.
b) Déterminer une équation de la droite de régression de "IDRS" selon "Densité".
$a = \dfrac{Cov(\text{Densité}, \text{IDRS})}{Var(\text{Densité})} = \dfrac{15949.50}{814.60} = 19.58.$
$b = 3272.1 - 19.58 \times 164 = 61.05$, d'où la droite de régression :
$\text{IDRS} = 19.58\text{Densité} + 61.05.$
c) Représenter graphiquement le nuage de points correspondant et la droite de régression.
Cf. figure 5.

2) En utilisant les données relatives à l'ensemble des départements (hors Ile de France) et d'autres variables explicatives, l'auteur propose le modèle suivant :

$$\text{IDRS} = 9.317X_1 - 19.919X_2 - 10.957X_3 + 59.081X_4 + 20.459X_5 + 490.184$$

où les variables X_1 à X_5 ont les significations suivantes :
X_1 : densité de médecins
X_2 : indicateur de mobilité de la clientèle des omnipraticiens
X_3 : indicateur de mobilité de la clientèle des spécialistes
X_4 : part des personnes âgées de 70 ans ou plus dans la population totale
X_5 : part des honoraires payés en tiers payant.

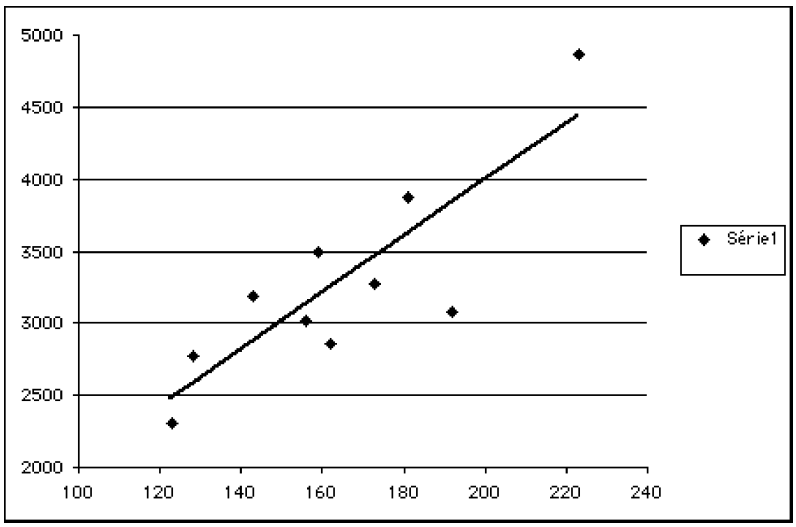

Figure 5: Nuage de points et droite de régression

a) Calculer l'IDRS estimé pour le département des Côtes d'Armor, pour lequel les variables précédentes ont pour valeurs respectives : $X_1 = 159.3$, $X_2 = 4.7$, $X_3 = -10.2$, $X_4 = 12.3$, $X_5 = 20.5$. Comparer IDRS estimé et l'IDRS observé : 3209.6.
$\widehat{IDRS} = 9.317 \times 159.3 - 19.919 \times 4.7 + 10.957 \times 10.2 + 59.081 \times 12.3 + 20.459 \times 20.5 + 490.184 = 3138.63$.
On constate que cette valeur est proche de l'IDRS observé (3209.60), les différences entre les départements (plus de 2500 entre l'Ain et la Haute-Corse) étant largement plus importantes que la différence constatée ($3209.60 - 3138.63 = 70.97$).
b) L'auteur indique que le coefficient de corrélation multiple vaut $R^2 = 0.98$. Quelle est la part de la variance de l'IDRS qui est "expliquée" par le modèle? Quel commentaire peut-on faire quant à la qualité du modèle proposé?
Autrement dit, 98 % de la variance de l'IDRS sont expliqués par le modèle. Il s'agit donc d'une modélisation d'excellente qualité.