INSTITUT NATIONAL DES SCIENCES APPLIQUÉES DE ROUEN

EC M8

Rapport de projet de M8

Titre du projet :

« Student performances »

Quels sont les principaux facteurs ayant une influence sur les résultats scolaires d'un élève ?
Quel est le degré d'influence de ces facteurs ?

Auteurs :
Pierre COIEFFEY
Gautier DARCHEN

21 juin 2015

Table des matières

Introduction

Dans le cadre de notre projet de M8, il nous fallait trouver des données à traiter statistiquement afin d'appliquer dans un cadre pratique ce que nous avions de manière théorique. Nous souhaitions initialement traiter des données portant sur le sport. Cependant, les données que nous trouvions ne satisfaisaient pas toutes les conditions que nous nous étions fixés : grands échantillons, diversité des variables, plusieurs types d'études statistiques à réaliser... Après plusieurs semaines de recherches et l'accord de nos chargés de TD – Messieurs Delporte, Canu et Rousselle – nous nous sommes mis d'accord sur un sujet assez original : l'étude de notes d'élèves portugais [1] dans différentes matières en fonction de multiples variables, liées à leur situation familiale, leurs habitudes de vie...

En premier lieu, nous nous sommes fixé comme objectif de cette étude statistique de répondre à notre problématique : Quels sont les principaux facteurs ayant une influence sur les résultats scolaires d'un élève ? Quel est le degré d'influence de ces facteurs ?

Afin de répondre à cette problématique, nous allons tout d'abord présenter les données que nous avons étudiées, à savoir l'ensemble des variables, ainsi que les modalités que chacune d'entre elles représente. Ensuite, nous allons concrétiser le traitement statistique des données de sorte à en sortir dégager des conclusions, notamment sur le plan mathématique. Enfin, notre étude se complètera au travers de tests. Nous allons ainsi réaliser quelques tests de Student et tests du χ^2.

1. Ces données sont disponibles à l'adresse `https://archive.ics.uci.edu/ml/datasets/Student+Performance`.

1 Présentation et récupération des données

Nous avons récupéré les données que nous avons choisi de traiter sur une librairie de données statistiques en ligne, à l'adresse `https://archive.ics.uci.edu/ml/datasets/Student+Performance`. Les données sont regroupées dans deux fichiers de type CSV (*comma-separated values*). Nous allons donc tout d'abord présenter les données, puis expliquer comment nous les avons récupérées afin de pouvoir les traiter de manière statistique.

1.1 Présentation des données

1.1.1 Présentation globale

Ces données décrivent les résultats d'étudiants de deux lycées au Portugal en fonction de certains critères. Les-dites données ont été rassemblées par le biais de questionnaires et de rapports réalisés par les écoles elles-mêmes. Parmi les critères, dont il nous faut déterminer les plus influents sur les résultats scolaires, figurent par exemple des caractéristiques démographiques, sociales propres aux différents étudiants ou liées aux écoles. Deux bases de données sont fournies dans les fichiers CSV, permettant d'étudier les performances des élèves dans deux matières différentes : les mathématiques dans le fichier `student-mat.csv` et le portugais – qui est leur langue maternelle – dans le fichier `student-pro.csv`. Les deux lycées étudiés sont GABRIEL PEREIRA et MOUSINHO DA SILVEIRA et cette enquête date de fin 2014. Rassemblons dans un tableau les données concernant les 15 premiers élèves du fichier `student-mat.csv`, sachant que d'un fichier à l'autre, l'architecture des données est vraisemblablement la même si ce n'est que les résultats obtenus sont ceux d'une matière scolaire différente.

TABLE 1 – Première partie des données sur les 15 premiers individus du fichier `student-mat.csv`

Elève	School	Sex	Age	Address	Famsize	Pstatus	Medu	Fedu	Mjob	Fjob	Reason
1	GP	F	18	U	GT3	A	4	4	at_home	teacher	course
2	GP	F	17	U	GT3	T	1	1	at_home	other	course
3	GP	F	15	U	LE3	T	1	1	at_home	other	other
4	GP	F	15	U	GT3	T	4	2	health	services	home
5	GP	F	16	U	GT3	T	3	3	other	other	home
6	GP	M	16	U	LE3	T	4	3	services	other	reputation
7	GP	M	16	U	LE3	T	2	2	other	other	home
8	GP	F	17	U	GT3	A	4	4	other	teacher	home
9	GP	M	15	U	LE3	A	3	2	services	other	home
10	GP	M	15	U	GT3	T	3	4	other	other	home
11	GP	F	15	U	GT3	T	4	4	teacher	health	reputation
12	GP	F	15	U	GT3	T	2	1	services	other	reputation
13	GP	M	15	U	LE3	T	4	4	health	services	course
14	GP	M	15	U	GT3	T	4	3	teacher	other	course
15	GP	M	15	U	GT3	A	2	2	other	other	home

TABLE 2 – Seconde partie des données sur les 15 premiers individus du fichier `student-mat.csv`

Elève	Guardian	Traveltime	Studytime	Failures	Schoolsup	Famsup	Paid	Activities	Nursery
1	mother	2	2	0	yes	no	no	no	yes
2	father	1	2	0	no	yes	no	no	no
3	mother	1	2	3	yes	no	yes	no	yes
4	mother	1	3	0	no	yes	yes	yes	yes
5	father	1	2	0	no	yes	yes	no	yes
6	mother	1	2	0	no	yes	yes	yes	yes
7	mother	1	2	0	no	no	no	no	yes
8	mother	2	2	0	yes	yes	no	no	yes
9	mother	1	2	0	no	yes	yes	no	yes
10	mother	1	2	0	no	yes	yes	yes	yes
11	mother	1	2	0	no	yes	yes	no	yes
12	father	3	3	0	no	yes	no	yes	yes
13	father	1	1	0	no	yes	yes	yes	yes
14	mother	2	2	0	no	yes	yes	no	yes
15	other	1	3	0	no	yes	no	no	yes

TABLE 3 – Troisième partie des données sur les 15 premiers individus du fichier `student-mat.csv`

Elève	Higher	Internet	Romantic	Famrel	Freetime	Goout	Dalc	Walc	Health	Absences
1	yes	no	no	4	3	4	1	1	3	6
2	yes	yes	no	5	3	3	1	1	3	4
3	yes	yes	no	4	3	2	2	3	3	10
4	yes	yes	yes	3	2	2	1	1	5	2
5	yes	no	no	4	3	2	1	2	5	4
6	yes	yes	no	5	4	2	1	2	5	10
7	yes	yes	no	4	4	4	1	1	3	0
8	yes	no	no	4	1	4	1	1	1	6
9	yes	yes	no	4	2	2	1	1	1	0
10	yes	yes	no	5	5	1	1	1	5	0
11	yes	yes	no	3	3	3	1	2	2	0
12	yes	yes	no	5	2	2	1	1	4	4
13	yes	yes	no	4	3	3	1	3	5	2
14	yes	yes	no	5	4	3	1	2	3	2
15	yes	yes	yes	4	5	2	1	1	3	0

TABLE 4 – Quatrième partie des données sur les 15 premiers individus du fichier `student-mat.csv`

Elève	G1	G2	G3
1	5	6	6
2	5	5	6
3	7	8	10
4	15	14	15
5	6	10	10
6	15	15	15
7	12	12	11
8	6	5	6
9	16	18	19
10	14	15	15
11	10	8	9
12	10	12	12
13	14	14	14
14	10	10	11
15	14	16	16

1.1.2 Description des variables

Les deux fichiers CSV contiennent les données de 33 variables : 32 sont explicatives et la 33ème variable représente les notes que pourraient obtenir les élèves au troisième trimestre et est donc la variable à expliquer. Décrivons chacune d'entre elles en détail.

1. ***School*** – L'école de l'individu concerné.
   Valeurs binaires :

- « GP » pour GABRIEL PEREIRA ;
- « MS » pour MOUSINHO DA SILVEIRA.

2. ***Sex*** – Le sexe de l'individu concerné.
   Valeurs binaires :
   - « F » pour féminin ;
   - « M » pour masculin.
3. ***Age*** – L'âge de l'étudiant concerné.
   Valeurs numériques allant de 15 à 22 ans.
4. ***Address*** – Le lieu de vie de l'individu concerné.
   Valeurs binaires :
   - « U » pour un mode de vie urbain ;
   - « R » pour un mode de vie rural.
5. ***Famsize*** – La taille de la famille de l'individu concerné.
   Valeurs binaires :
   - « LE3 » pour *less or equal to 3* ;
   - « GT3 » pour *greater than 3*.
6. ***Pstatus*** – Le statut de cohabitation des parents de l'individu concerné.
   Valeurs binaires :
   - « T » pour *living together* ;
   - « A » pour *apart*).
7. ***Medu*** – Le niveau d'éducation de la mère de l'individu concerné.
   Valeurs numériques :
   - 0 si la mère n'est pas allée à l'école ;
   - 1 si elle a passé moins de 4 ans à l'école (primaire) ;
   - 2 si elle a passé entre 5 et 9 ans à l'école (fin primaire et collège) ;
   - 3 si la mère a eu une éducation secondaire [2] (lycée) ;
   - 4 si elle est allée dans des écoles d'enseignement supérieur.
8. ***Fedu*** – Le niveau d'éducation du père de l'individu concerné.
   Valeurs numériques :
   - 0 si le père n'est pas allé à l'école ;
   - 1 si le père a passé moins de 4 ans à l'école (primaire) ;
   - 2 si le père a passé entre 5 et 9 ans à l'école (fin primaire et collège) ;
   - 3 si le père a eu une éducation secondaire (lycée) ;
   - 4 si il est allé dans des écoles d'enseignement supérieur.
9. ***Mjob*** – Le métier de la mère de l'individu concerné.
   Valeurs nominales :
   - « *teacher* » si elle travaille dans l'enseignement ;
   - « *health* » si elle travaille dans le domaine de la santé ;
   - « *services* » si elle travaille dans les services publics (par exemple dans l'administratif ou la police) ;
   - « *at_home* » si elle est femme au foyer ;
   - « *other* » si son métier ne rentre dans aucune de ces catégories.
10. ***Fjob*** – Le métier du père de l'individu concerné.
    Valeurs nominales :
    - « *teacher* » si il travaille dans l'enseignement ;
    - « *health* » si il travaille dans le domaine de la santé ;

2. Dans le système scolaire portugais, il s'agit de l'équivalent du Lycée en France.

- « *services* » si il travaille dans les services publics (par exemple dans l'administratif ou la police) ;
- « *at_home* » si il est homme au foyer ;
- « *other* » si son métier ne rentre dans aucune de ces catégories.

11. ***Reason*** – La raison pour laquelle l'individu concerné a choisi l'école dans laquelle il étudie.
Valeurs nominales :
- « *home* » si ce choix est dû à la proximité de l'école avec la maison ;
- « *reputation* » si c'est grâce à la réputation de l'école ;
- « *course* » si ce choix est dû à l'intérêt pour les matières enseignées par l'école ;
- « *other* » si le choix est dû à un autre critère.

12. ***Guardian*** – Le responsable légal ou personne qui s'occupe de l'individu concerné.
Valeurs nominales :
- « *mother* » si le tutrice est la mère ;
- « *father* » si le tuteur est le père ;
- « *other* » s'il s'agit d'une autre personne.

13. ***Traveltime*** – Le temps de trajet moyen entre l'école et le domicile de l'étudiant concerné.
Valeurs numériques :
- 1 si le trajet dure en moyenne moins de 15 minutes ;
- 2 si le trajet dure en moyenne entre 15 et 30 minutes ;
- 3 si le trajet dure en moyenne entre 30 minutes et une heure ;
- 4 si le trajet dure en moyenne plus d'une heure.

14. ***Studytime*** – Le temps que passe l'élève concerné chaque semaine sur son travail personnel (révisions).
Valeurs numériques :
- 1 s'il étudie en moyenne moins de 2 heures par semaine ;
- 2 s'il étudie en moyenne entre 2 et 5 heures par semaine ;
- 3 s'il étudie en moyenne entre 5 et 10 heures par semaine ;
- 4 s'il étudie en moyenne plus de 10 heures par semaine.

15. ***Failures*** – Le nombre de classes que l'individu a redoublé par le passé.
Valeurs numériques :
- n, avec $1 \leq n < 3$ si l'élève a redoublé n classe(s) ;
- 4 si l'élève a redoublé plus de 3 fois.

16. ***Schoolsup*** – Un booléen décrivant si l'individu dispose d'un soutien éducatif supplémentaire (dans l'école).
Valeurs binaires (booléen) :
- *yes* si l'individu dispose d'un soutien au sein de l'école ;
- *no* sinon.

17. ***Famsup*** – Un booléen décrivant si l'individu dispose d'un soutien pédagogique familial.
Valeurs binaires (booléen) :
- *yes* si l'individu dispose d'un soutien pédagogique familial ;
- *no* sinon.

18. ***Paid*** – Un booléen décrivant si l'individu prend des cours particuliers (payants) dans l'une des deux matières étudiées (mathématiques ou portugais).
Valeurs binaires (booléen) :
- *yes* si l'individu prend des cours particuliers payants ;
- *no* sinon.

19. ***Activities*** – Un booléen décrivant si l'individu pratique des activités extra-scolaires.
Valeurs binaires (booléen) :

- *yes* si l'individu pratique au moins une activité extra-scolaire ;
- *no* sinon.

20. ***Nursery*** – Un booléen décrivant si l'individu est allé à l'école maternelle.
Valeurs binaires (booléen) :
- *yes* si l'individu est allé à la maternelle ;
- *no* sinon.

21. ***Higher*** – Un booléen décrivant si l'individu a pour intention, au moment de l'enquête, de faire des études supérieures.
Valeurs binaires (booléen) :
- *yes* si l'individu a pour intention de faire des études supérieures ;
- *no* sinon.

22. ***Internet*** – Un booléen décrivant si l'individu a un accès à Internet à son domicile.
Valeurs binaires (booléen) :
- *yes* si possède un accès à Internet chez lui ;
- *no* sinon.

23. ***Romantic*** – Un booléen décrivant si l'individu est en couple ou non.
Valeurs binaires (booléen) :
- *yes* si l'individu est en couple au moment de l'enquête ;
- *no* sinon.

24. ***Famrel*** – Une valeur décrivant la qualité de la relation entre l'individu et les membres de sa famille.
Valeurs numériques (de 1 à 5) :
- 1 si les relations sont très mauvaises ;

  ⋮
- 5 si ces relations sont excellentes.

25. ***Freetime*** – Une valeur décrivant le temps libre dont dispose l'étudiant après les cours.
Valeurs numériques (de 1 à 5) :
- 1 si l'étudiant a très peu de temps libre après les cours ;

  ⋮
- 5 si l'étudiant a beaucoup de temps libre après les cours.

26. ***Goout*** – Une valeur numérique décrivant le temps que l'élève passe avec ses amis hors de chez lui.
Valeurs numériques (de 1 à 5) :
- 1 si l'élève sort très peu avec ses amis ;

  ⋮
- 5 si l'élève sort beaucoup avec ses amis.

27. ***Dalc*** – Une valeur numérique représentant la quantité d'alcool que consomme l'étudiant quotidiennement.
Valeurs numériques (de 1 à 5) :
- 1 si l'étudiant consomme très peu d'alcool au quotidien ;

  ⋮
- 5 si l'étudiant consomme beaucoup d'alcool au quotidien.

28. ***Walc*** – Une valeur numérique représentant la quantité d'alcool que consomme l'étudiant chaque semaine.
Valeurs numériques (de 1 à 5) :
- 1 si l'étudiant consomme très peu d'alcool par semaine ;

  ⋮

- 5 si l'étudiant consomme beaucoup d'alcool par semaine.

29. ***Health*** – Une valeur numérique représentant l'état de santé de l'étudiant au moment de l'enquête (santé physique et/ou psychologique).
Valeurs numériques (de 1 à 5) :
   - 1 si l'étudiant était en très mauvaise santé au moment de l'enquête ;
   - ⋮
   - 5 si l'étudiant était en excellente santé au moment de l'enquête.
30. ***Absences*** – Le nombre de jours durant lesquels l'élève a été absent.
Valeurs numériques (de 0 à 93) :
   - 0 si l'élève n'a jamais été absent durant l'année de l'enquête ;
   - ⋮
   - 93 si l'élève a été absent 93 fois pendant l'année (il s'agit de la valeur maximale d'absentéisme dans cette enquête).
31. ***G1*** – La note finale de l'étudiant à la fin du premier trimestre dans la matière en question (mathématiques ou portugais).
Valeurs numériques (de 0 à 20) :
   - 0 : note minimale ;
   - ⋮
   - 20 : note maximale.
32. ***G2*** – La note finale de l'étudiant à la fin du second trimestre dans la matière en question (mathématiques ou portugais).
Valeurs numériques (de 0 à 20) :
   - 0 : note minimale ;
   - ⋮
   - 20 : note maximale.
33. ***G3*** – La note finale de l'étudiant à la fin du troisième trimestre dans la matière en question (mathématiques ou portugais). Il s'agit de l' « *output target* », autrement dit, la variable à expliquer.
Valeurs numériques (de 0 à 20) :
   - 0 : note minimale ;
   - ⋮
   - 20 : note maximale.

1.2 Récupération des données

Les données sont stockées dans des fichiers CSV. Nous décrirons ici la manière dont nous avons extrait les données du fichier `student-mat.csv`, sachant que nous avons procédé exactement de la même façon pour le fichier `student-por.csv`. Le code de cette extraction est réalisé dans le fichier `DonneesProjetM8.m`.

La première étape consiste à charger une matrice contenant les données. Pour les importer et pouvoir les exploiter avec le logiciel *Matlab*, nous avons donc tout d'abord dû utiliser la fonction `csvimport`. Ainsi, la première ligne de code de cette extraction est :

Code 1 – Extraction des données du fichier CSV dans une matrice de `cell`

```
2 dataMat=csvimport('student-mat.csv');
```

Nous avons alors la matrice `dataMat` qui est de type `cell`. Il faut donc convertir ces valeurs afin de pouvoir les exploiter. Cependant, il ne faut pas oublier que les variables ont des types différents. En effet, comme nous l'avons vu précédemment, certaines variables admettent des modalités nominales, d'autres des modalités numériques...

Ainsi, nous convertissons les données de type `cell` dont nous voulons obtenir des valeurs numériques. Toutes les variables obtenues par le code suivant seront alors de type `double` et nous pourrons les exploiter.

Code 2 – Transformation de `cell` en `double`

```
4  % Donnees numeriques
5  Mat_Age=cell2mat(dataMat(2:end,3));
6  Mat_Medu=cell2mat(dataMat(2:end,7));
7  Mat_Fedu=cell2mat(dataMat(2:end,8));
8  Mat_Traveltime=cell2mat(dataMat(2:end,13));
9  Mat_Studytime=cell2mat(dataMat(2:end,14));
10 Mat_Failures=cell2mat(dataMat(2:end,15));
11 Mat_Famrel=cell2mat(dataMat(2:end,24));
12 Mat_Freetime=cell2mat(dataMat(2:end,25));
13 Mat_Goout=cell2mat(dataMat(2:end,26));
14 Mat_Dalc=cell2mat(dataMat(2:end,27));
15 Mat_Walc=cell2mat(dataMat(2:end,28));
16 Mat_Health=cell2mat(dataMat(2:end,29));
17 Mat_Absences=cell2mat(dataMat(2:end,30));
18 Mat_G1=cell2mat(dataMat(2:end,31));
19 Mat_G2=cell2mat(dataMat(2:end,32));
20 Mat_G3=cell2mat(dataMat(2:end,33));
```

Par ailleurs, nous voulons que certaines variables, comme celle indiquant le métier de la mère d'un élève par exemple, soient de type char puisqu'elles représentent des chaînes de caractères. C'est alors tout l'intérêt du code qui suit.

Code 3 – Transformation de `cell` en `char`

```
22 % Donnees en chaines de caracteres
23 Mat_School=char(dataMat(2:end,1));
24 Mat_Sex=char(dataMat(2:end,2));
25 Mat_Address=char(dataMat(2:end,4));
26 Mat_Famsize=char(dataMat(2:end,5));
27 Mat_Pstatus=char(dataMat(2:end,6));
28 Mat_Mjob=char(dataMat(2:end,9));
29 Mat_Fjob=char(dataMat(2:end,10));
30 Mat_Reason=char(dataMat(2:end,11));
31 Mat_Guardian=char(dataMat(2:end,12));
32 Mat_Schoolsup=char(dataMat(2:end,16));
33 Mat_Famsup=char(dataMat(2:end,17));
34 Mat_Paid=char(dataMat(2:end,18));
35 Mat_Activities=char(dataMat(2:end,19));
36 Mat_Nursery=char(dataMat(2:end,20));
37 Mat_Higher=char(dataMat(2:end,21));
38 Mat_Internet=char(dataMat(2:end,22));
39 Mat_Romantic=char(dataMat(2:end,23));
```

A ce stade, nous réalisons exactement les mêmes transformations pour les données contenues dans le fichier `student-por.csv`.

Ensuite, il faut ranger les données de type `char` en valeurs numériques. On associe alors une va-

leur numérique à une modalité en chaîne de caractères. Les différentes valeurs numériques associées aux modalités d'une variable sont indiquées dans le code qui suit.

Code 4 – Extrait du rangement des modalités de type `char` avec des valeurs numériques de type `double`

```
%% Rangement modalites (chaines) fichier 'student_mat.csv'

%Rangement modalites Mat_School (0=MS / 1=GP)
TabModMat_School=ones(length(Mat_School),1);
for i=1:length(Mat_School)
    ind=[];
    if (Mat_School(i,:)=='MS')
       ind=[ind ;i];
       TabModMat_School(ind,1)=0;
    end
end

%Rangement modalites Mat_Sex (0=F / 1=M)
TabModMat_Sex=ones(length(Mat_Sex),1);
for i=1:length(Mat_Sex)
    ind=[];
    if (Mat_Sex(i,:)=='F')
       ind=[ind ;i];
       TabModMat_Sex(ind,1)=0;
    end
end

%Rangement modalites Mat_Address (0=U / 1=R)
TabModMat_Address=ones(length(Mat_Address),1);
for i=1:length(Mat_Address)
    ind=[];
    if (Mat_Address(i,:)=='U')
       ind=[ind ;i];
       TabModMat_Address(ind,1)=0;
    end
end

%Rangement modalites Mat_Famsize (0=LE3 / 1=GT3)
TabModMat_Famsize=ones(length(Mat_Famsize),1);
for i=1:length(Mat_Famsize)
    ind=[];
    if (Mat_Famsize(i,:)=='LE3')
       ind=[ind ;i];
       TabModMat_Famsize(ind,1)=0;
    end
end

%Rangement modalites Mat_Pstatus (0=A / 1=T)
TabModMat_Pstatus=ones(length(Mat_Pstatus),1);
for i=1:length(Mat_Pstatus)
    ind=[];
    if (Mat_Pstatus(i,:)=='A')
       ind=[ind ;i];
       TabModMat_Pstatus(ind,1)=0;
    end
end

%Rangement modalites Mat_Mjob (0=at_home / 1=health / 2=other / 3=services / 4=teacher)
TabModMat_Mjob=ones(length(Mat_Mjob),1);
for i=1:length(Mat_Mjob)
    ind=[];
```

```
143      if (Mat_Mjob(i,:)=='at_home ')
144         ind=[ind ;i];
145         TabModMat_Mjob(ind,1)=0;
146      end
147      if (Mat_Mjob(i,:)=='other   ')
148         ind=[ind ;i];
149         TabModMat_Mjob(ind,1)=2;
150      end
151      if (Mat_Mjob(i,:)=='services')
152         ind=[ind ;i];
153         TabModMat_Mjob(ind,1)=3;
154      end
155      if (Mat_Mjob(i,:)=='teacher ')
156         ind=[ind ;i];
157         TabModMat_Mjob(ind,1)=4;
158      end
159  end
160
161  %Et ainsi de suite pour toutes les autres variables et l'autre fichier
```

A ce stade, les données sont rangées et il nous reste à les traiter statistiquement. Ce traitement se fera dans un autre fichier `Traitement.m` qui fera appel au fichier `DonneesProjetM8.m` que nous venons d'analyser.

2 Traitement des données

2.1 Etude des notes finales des élèves

Dans un premier temps, nous avons décidé de traiter la variable des notes finales des élèves toute seule pour nous familiariser un peu plus avec le sujet.

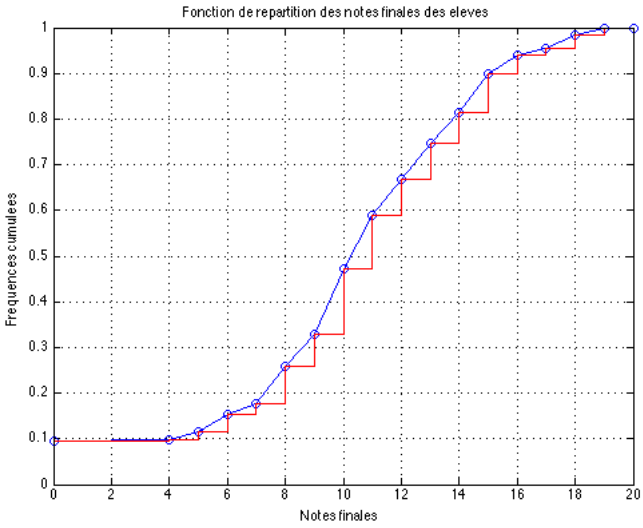

FIGURE 1 – Fonction de répartition des notes finales des élèves

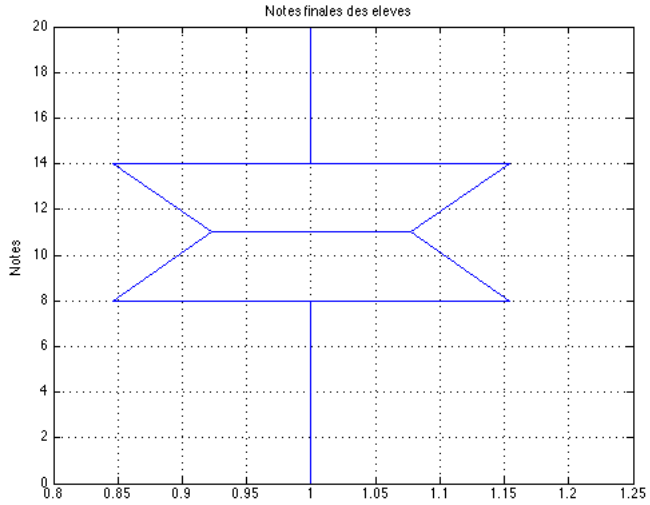

FIGURE 2 – Boîte à moustache des notes finales des élèves

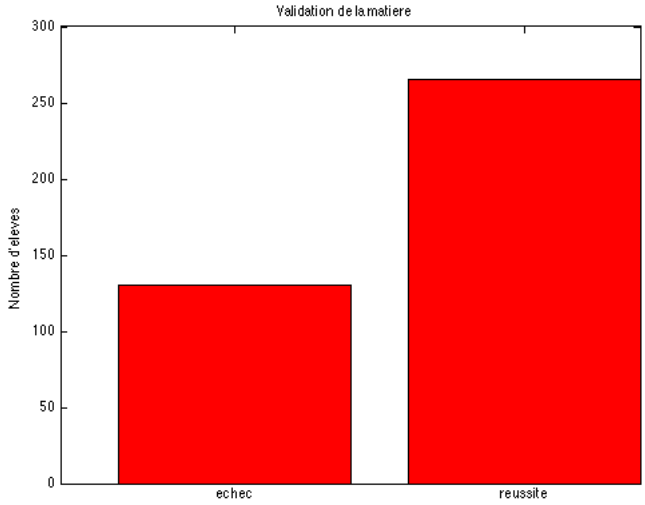

FIGURE 3 – Histogramme des élèves validant la matière Mathématiques

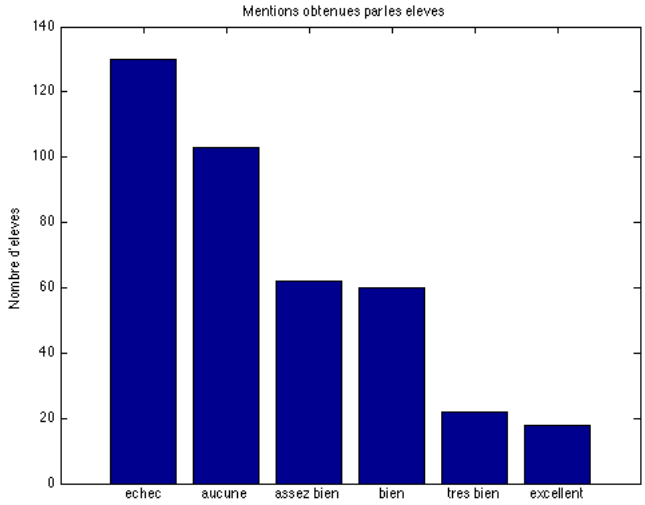

FIGURE 4 – Histogramme des mentions obtenues par les élèves

On remarque donc grâce à ces quelques graphiques que le taux d'échec des élèves dans cette matière paraît plutôt élevé avec 32,91% d'élèves ne validant pas leur année. Cependant ce chiffre doit être relativisé étant donné qu'il ne concerne qu'une seule matière. Les principaux indicateurs de cet échantillon de notes sont quant à eux plutôt « normaux » avec une moyenne de 10.42, une médiane de 11, un premier quartile égal à 8 et un troisième égal à 14, pour une distance inter-quartiles de 6, et avec des notes occupant toute l'échelle de notation. En effet, la notation adoptée tend à suivre une courbe de GAUSS c'est-à-dire qu'il y a beaucoup de notes moyennes pour très peu de notes extrêmes comme le montrent les mentions obtenues par les élèves.

Après avoir étudié ces notes finales, il est à présent temps de revenir à la problématique de notre sujet et de nous intéresser aux variables influençant cette fameuse note finale. Et pour ce faire, nous allons réaliser une Analyse en Composantes Principales (ACP) qui nous permettra d'avoir une vue d'ensemble sur toutes les variables pouvant influencer positivement ou négativement les notes finales des élèves.

2.2 Analyse en Composantes Principales

La première chose à faire dans une ACP est de centrer et réduire les données pour éviter que des variables possédant des modalités très élevées prennent le pas sur d'autres ayant des modalités très faibles.

```
62 Xc_Mat=(X_Mat-moyenne);% matrice centree
63 Xn_Mat=Xc_Mat./ecart_type;% matrice centree et reduite
```

La deuxième étape, quant à elle, consiste à calculer les vecteurs et valeurs propres de cette matrice.

```
65 [V D]=eig(Xn_Mat'*Xn_Mat);% calcul des vecteurs et valeurs propres
66 lambda=diag(D);% valeurs propres
```

Une fois cette étape réalisée, il faut calculer le pourcentage d'information porté par chacune de nos valeurs propres. Malheureusement, comme nous avons de très nombreuses variables, les pourcentages portés par chacune de nos valeurs propres vont être assez faibles et par conséquent notre ACP ne va pas pas être en mesure de représenter plus d'un quart des informations. Cependant, cette analyse va quand même nous permettre d'avoir une vue d'ensemble des corrélations entre variables et va nous permettre d'orienter la suite de notre traitement de données.

TABLE 5 – Valeurs propres et pourcentages d'information

λ	**Pourcentage d'information**
30.572	0.23514
64.432	0.49555
109.57	0.84272
118.98	0.91509
156.61	1.2045
180.52	1.3884
194.99	1.4997
210.82	1.6214
220.43	1.6954
234.49	1.8035
247.9	1.9067
252.05	1.9386
277.56	2.1348
282.37	2.1717
300.06	2.3078
314.17	2.4163
315.76	2.4286
331.84	2.5522
358.1	2.7542
376.58	2.8963
381.59	2.9349
401.41	3.0873
422.88	3.2525
436.22	3.355
471.01	3.6226
489.44	3.7643
550.89	4.2369
564.17	4.3391
600.73	4.6203
701.2	5.393
878.27	6.7548
1000.4	7.6945
1526	11.736

Enfin, la dernière étape est de visualiser les variables en les projetant sur les vecteurs propres associés aux plus grandes valeurs propres. Ici nous avons choisi les 2 plus grandes après avoir conclu que l'ajout de la troisième n'apportait rien de plus si ce n'est une visualisation plus complexe. De même, nous avons

remarqué que la visualisation des individus n'apportait rien à l'analyse. En effet, le fait d'avoir un nombre très important d'individus dans notre base de données rendait la visualisation brouillonne et ne permettait pas de tirer un quelconque enseignement de cette analyse.

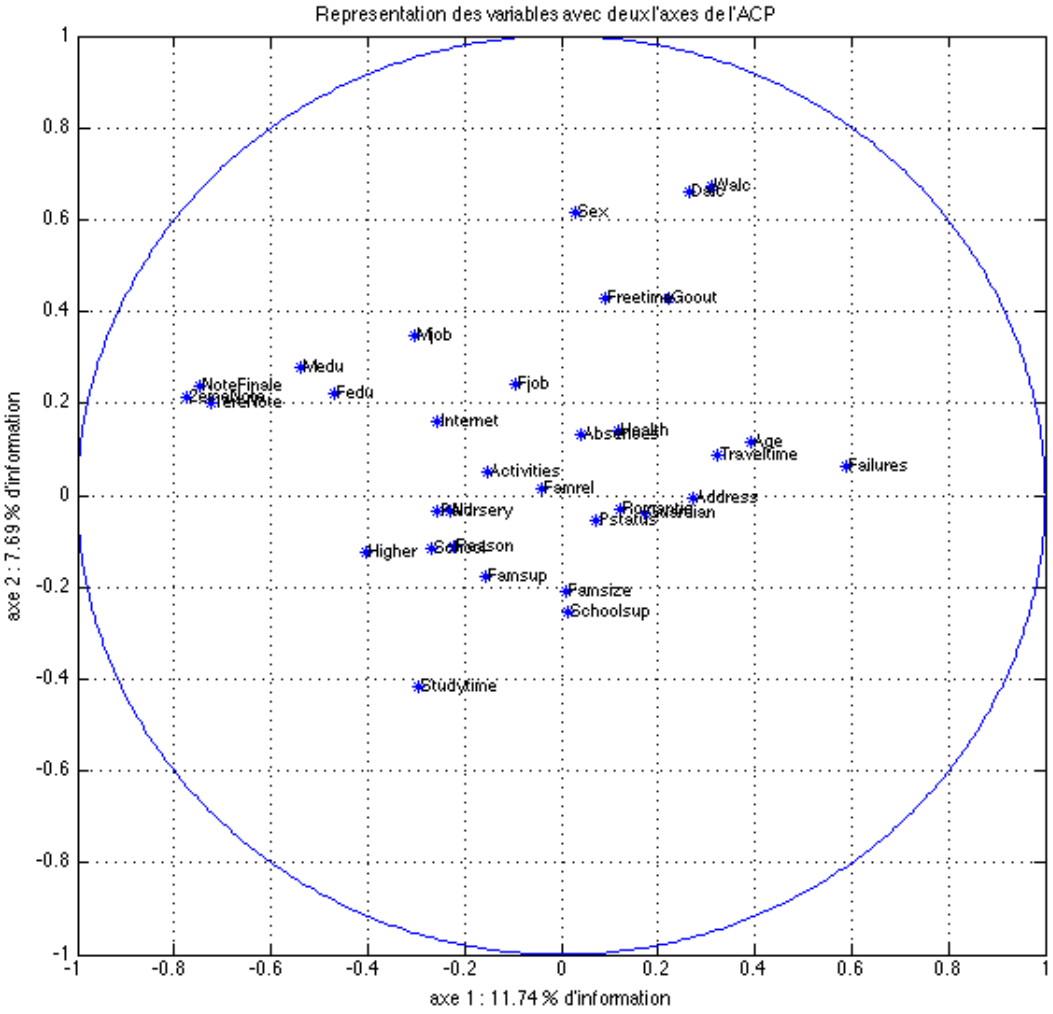

FIGURE 5 – Résultat de l'ACP après projection sur les composantes principales

On peut voir sur ce graphique que les première et deuxième notes sont très fortement corrélées positivement avec la note finale. Ceci est logique étant donné que ces trois variables sont du même type et cela signifie que si les premières et deuxième notes sont élevées, la note finale le sera également (et inversement). Nous essayerons donc, sachant cela, de prédire la note finale à l'aide des deux premières. Concernant les autres variables pouvant avoir une influence sur la note finale, nous retrouvons `Medu` et `Fedu` corrélées positivement avec la note finale mais également `Failures` et `Age` corrélées négativement. Tout cela signifie que plus l'éducation des parents est forte et plus la note finale sera élevée (et inversement) contrairement aux deux autres variables qui indiquent que plus les absences et les âges sont élevés et plus la note finale sera faible (et inversement). Bien-entendu, ces conclusions ne sont qu'intermédiaires et elles demanderont à être vérifiées par la suite pour voir si oui ou non ces variables ont une influence sur la note finale.

2.3 Régression linéaire

But : Peut-on prédire les notes finales des élèves à l'aide de leurs deux premières notes ?

2.3.1 Première régression

Régression simple : Peut-on prédire les notes finales des élèves à l'aide de leur première note ?

— Variable explicative : `Mat_G1` c'est-à-dire celle contenant les premières notes des élèves.
— Variable à expliquer : `Mat_G3` c'est-à-dire celle contenant les notes finales des élèves.

Une fois la régression réalisée on obtient les paramètres suivants : $a = 1,1063$; $b = -1,6528$ et un coefficient de détermination $R^2 = 0,6424$.

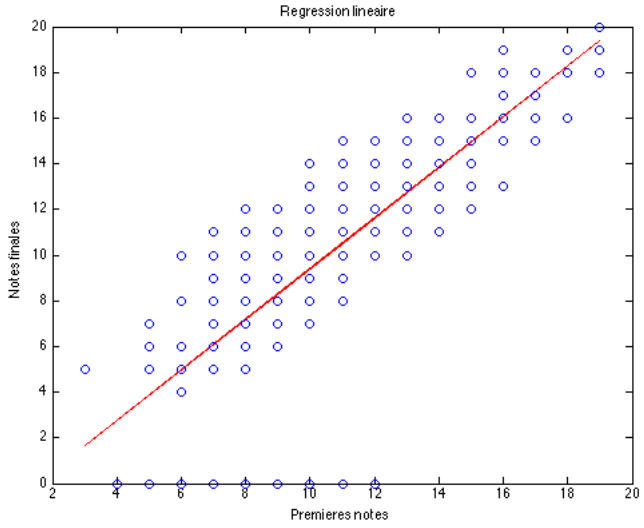

FIGURE 6 – Régression linéaire de la troisième note en fonction de la première

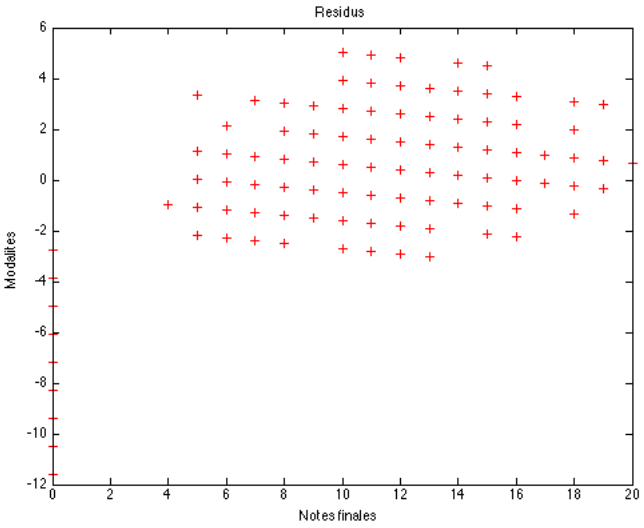

FIGURE 7 – Résidus selon les notes finales observées

Cependant, en regardant le graphique de la régression, celui des résidus, et celui des contributions, on remarque la présence de points aberrants correspondants tous aux notes égales à zéro. Ces points ne sont pas aberrants en soi étant donné que le zéro fait partie de l'échelle de notation. Cependant ces points ont

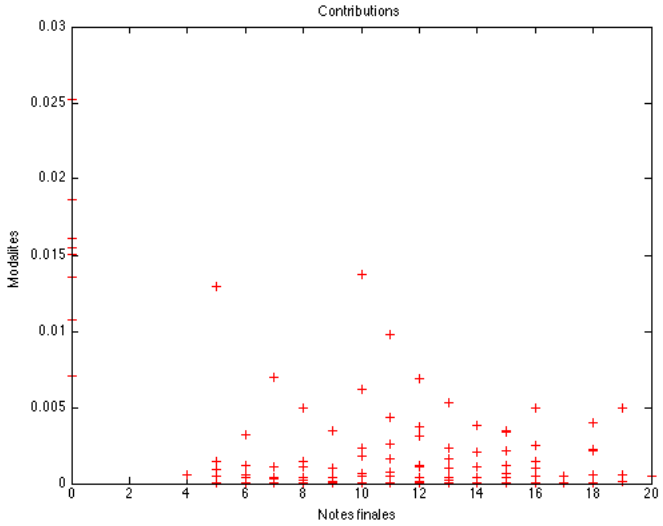

FIGURE 8 – Contributions selon les notes finales

tendance à faire chuter la qualité de la régression. C'est pourquoi après avoir réalisé une nouvelle régression avec l'absence de ces notes nous obtenons de nouveaux paramètres et un coefficient de détermination bien meilleur : $a = 0,8883$; $b = 1,5134$ et $R^2 = 0,7953$.

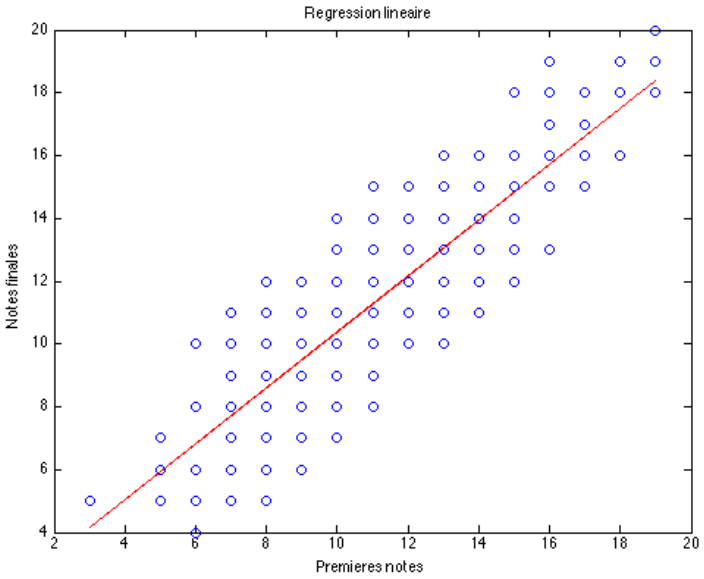

FIGURE 9 – Nouvelle régression linéaire de la troisième note en fonction de la première note, sans les points aberrants

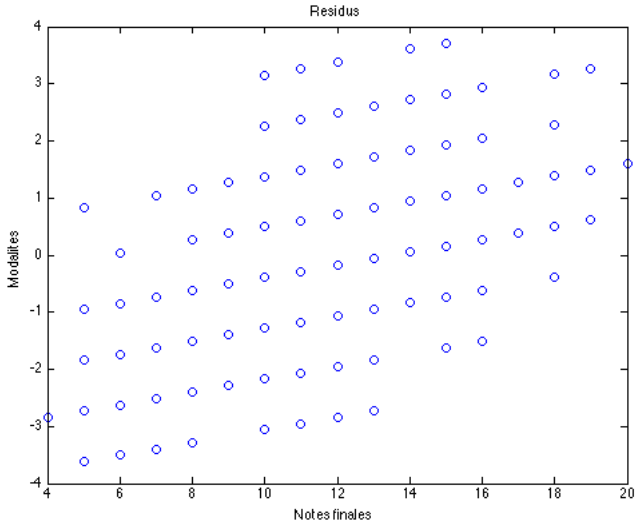

FIGURE 10 – Résidus selon les notes finales observées, sans les points aberrants

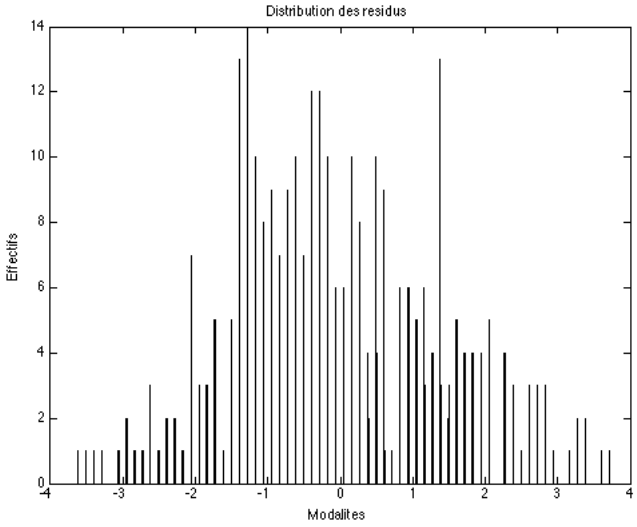

FIGURE 11 – Distribution des résidus, sans les points aberrants

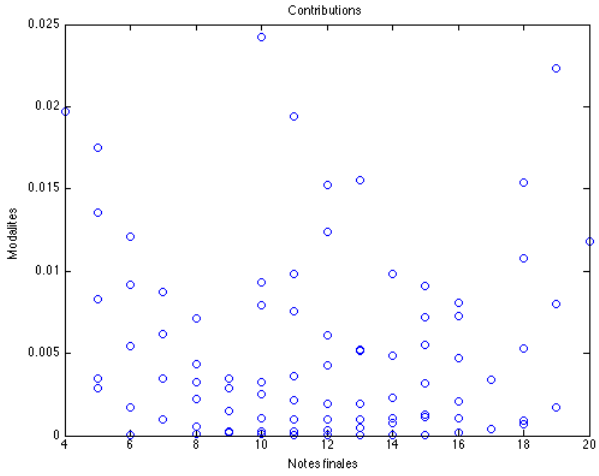

FIGURE 12 – Contributions selon les notes finales, sans les points aberrants

Avec l'appui de ces quelques graphiques et du coefficient de détermination nous nous apercevons que le modèle linéaire est le bon. En effet, R^2 est plutôt proche de 1, les résidus et les contributions sont homogènes, il n'y a donc plus de points aberrants et la distribution des résidus est normale (loi gaussienne). Tous ces facteurs tendent à montrer que la prédiction de la note finale, à l'aide de la première note des élèves, est tout à fait possible. Voyons à présent s'il en est de même avec la deuxième note.

2.3.2 Seconde régression

Régression simple : Peut-on prédire les notes finales des élèves à l'aide de leur deuxième note ?

— Variable explicative : `Mat_G2` c'est-à-dire celle contenant les deuxièmes notes des élèves.
— Variable à expliquer : `Mat_G3` c'est-à-dire celle contenant les notes finales des élèves.

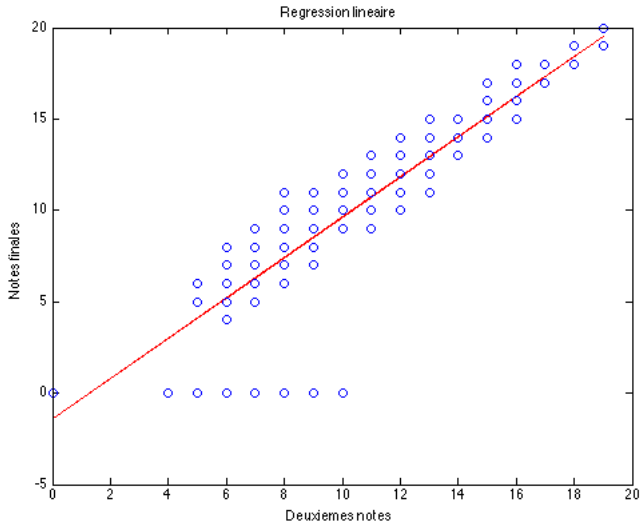

FIGURE 13 – Régression linéaire de la troisième note en fonction de la seconde

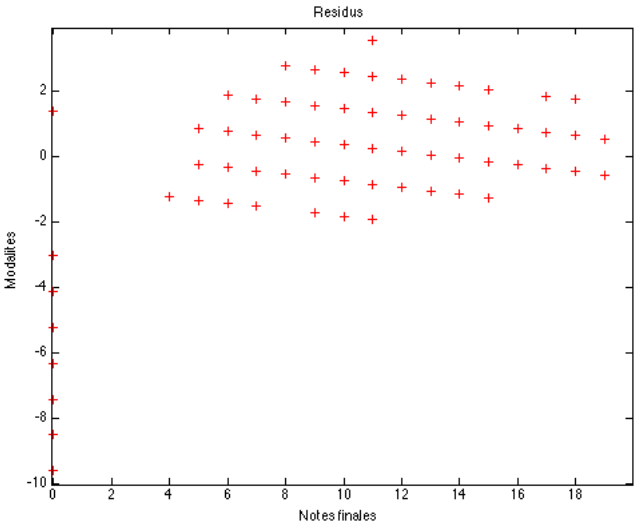

FIGURE 14 – Résidus selon les notes finales observées

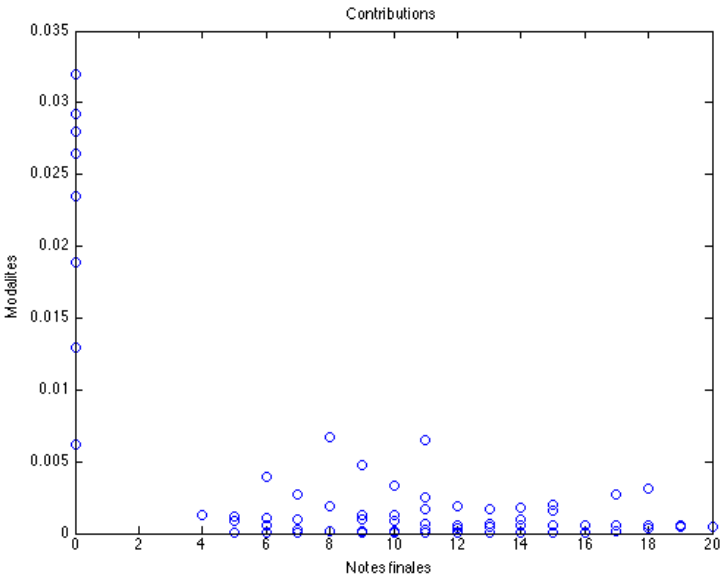

FIGURE 15 – Contributions selon les notes finales

Une fois la régression réalisée nous nous apercevons une nouvelle fois que les notes égales à 0 viennent fausser notre régression. Cependant les paramètres et le coefficient de détermination restent tout à fait corrects : $a = 1,1021$; $b = -1,3928$ et $R^2 = 0,8188$. En effet on s'aperçoit que cette régression est de meilleure facture que la précédente, même en y laissant les notes égales à 0. Voyons à présent la qualité de cette régression en y enlevant ces notes.

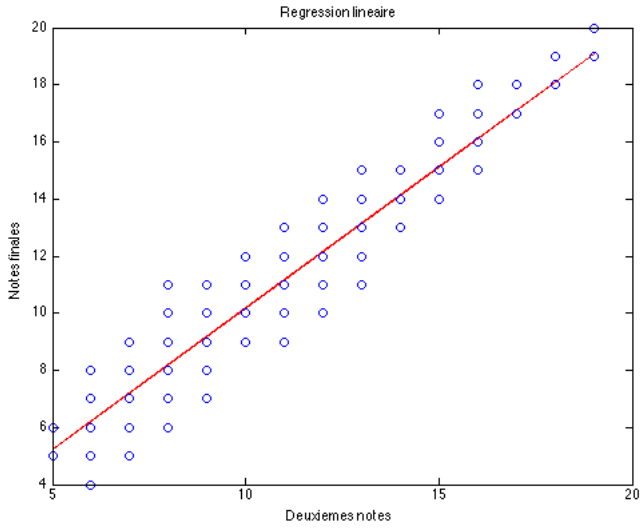

FIGURE 16 – Nouvelle régression linéaire de la troisième note en fonction de la seconde note, sans les points aberrants

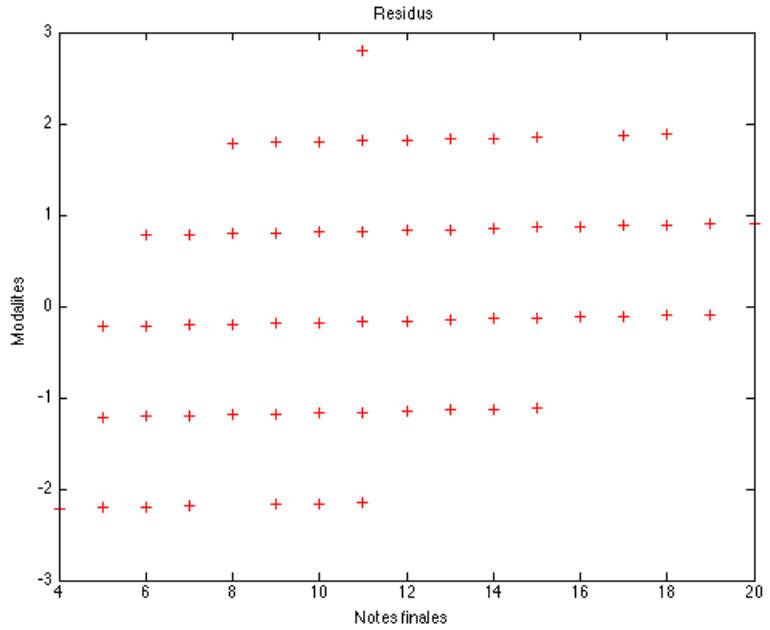

FIGURE 17 – Résidus selon les notes finales, sans les points aberrants

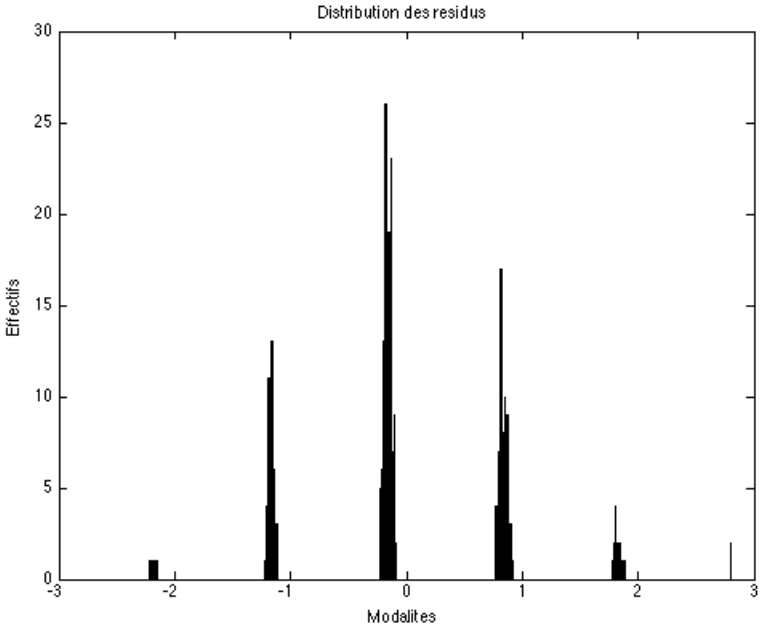

FIGURE 18 – Distribution des résidus selon les notes finales observées, sans les points aberrants

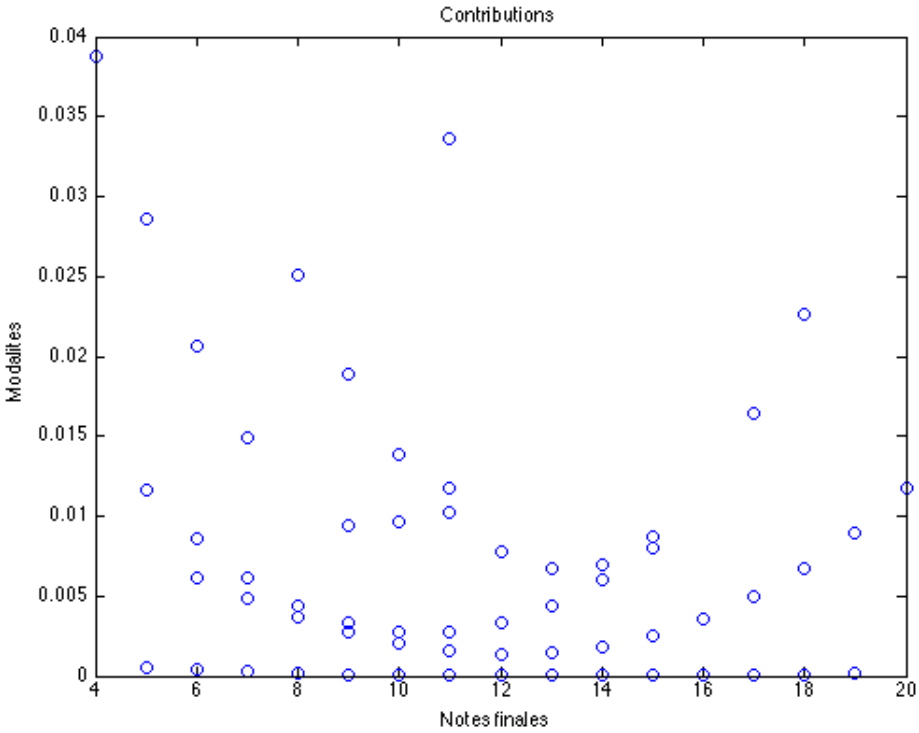

FIGURE 19 – Contributions selon les notes finales, sans les points aberrants

On obtient donc de nouveaux paramètres et, comme on pouvait s'y attendre, un coefficient de détermination meilleur : $a = 0,9903$; $b = 0,2753$ et $R^2 = 0,9323$. De plus, l'étude de ces graphiques nous montre, comme lors de la régression précédente, que le modèle linéaire est excellent et que nous sommes capables de prédire la note finale, à l'aide de la deuxième note des élèves. Cependant, on peut ajouter que cette régression est bien meilleure que la précédente. En effet, le coefficient de détermination est passé de $0,7953$ à $0,9323$ ce qui signifie que la deuxième note des élèves est plus représentative de la note finale que la première. Il vaut donc mieux utiliser cette seconde note pour prédire la note finale. Mais serait t-il possible d'avoir une meilleure prédiction en utilisant à la fois la première et la deuxième note des élèves ?

2.3.3 Troisième régression

Régression multiple : Peut-on prédire les notes finales des élèves à l'aide de leurs premières et deuxième notes ?

— Variables explicatives : `Mat_G1` c'est-à-dire celle contenant la première note de chaque élève et `Mat_G2` c'est-à-dire celle contenant la deuxième note de chaque élève.
— Variable à expliquer : `Mat_G3` c'est-à-dire celle contenant la note finale des élèves.

Une fois cette régression réalisée on obtient les paramètres et le coefficient de détermination suivants : $a_1 = 0,1533$; $a2 = 0,9869$; $b = -1,83$ et $R^2 = 0,8222$.

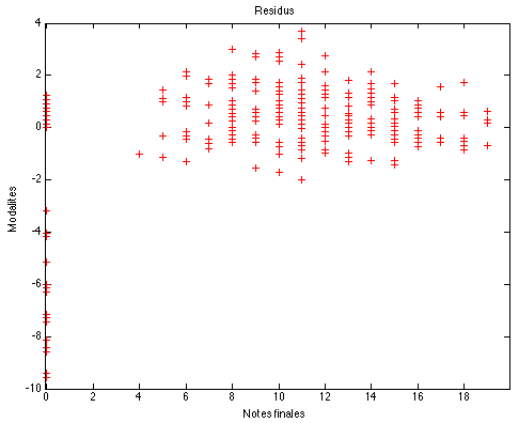

FIGURE 20 – Résidus selon les notes finales

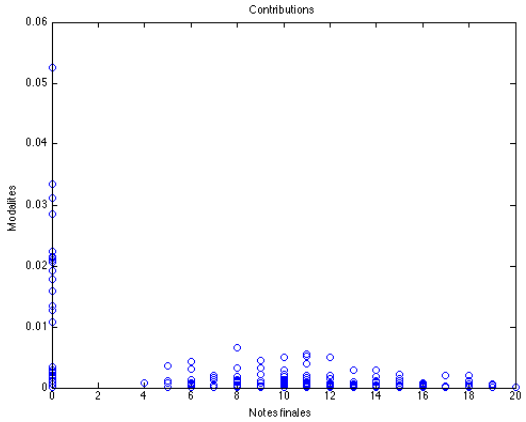

FIGURE 21 – Contribution selon les notes finales

On remarque une nouvelle fois la présence de ces notes égales à 0 qui viennent diminuer la qualité de la régression. Malgré cela la régression reste assez correcte avec un coefficient de détermination plutôt proche de 1. Cette régression est même de meilleure qualité que la première mais elle ne parvient pas à concurrencer la deuxième (lorsque l'on a retiré les points aberrants de la seconde régression). Voyons maintenant le résultat en enlevant ces points aberrants.

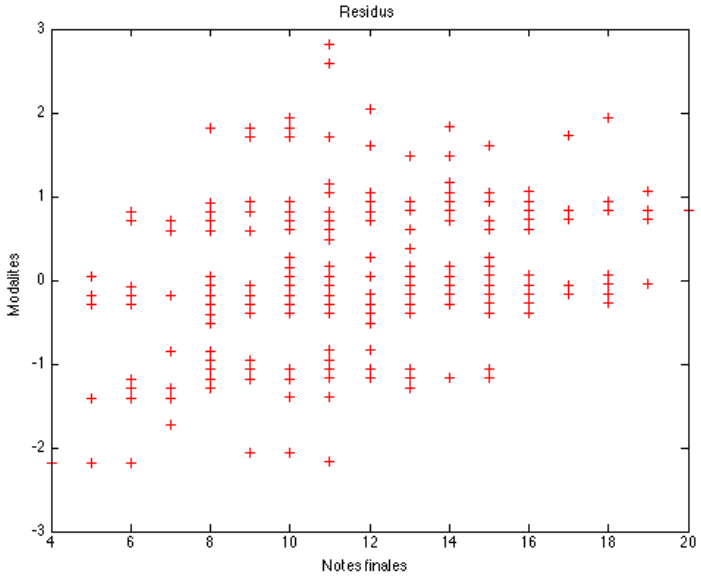

FIGURE 22 – Résidus selon les notes finales, sans les points aberrants

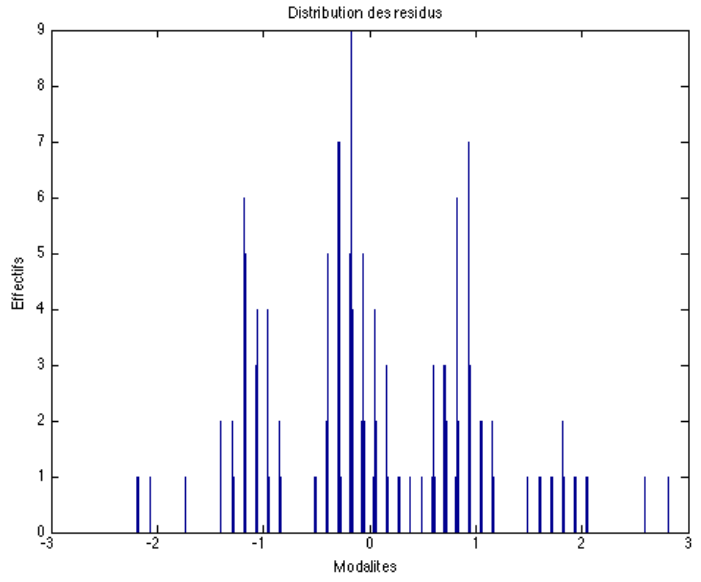

FIGURE 23 – Distribution des résidus selon les notes finales, sans les points aberrants

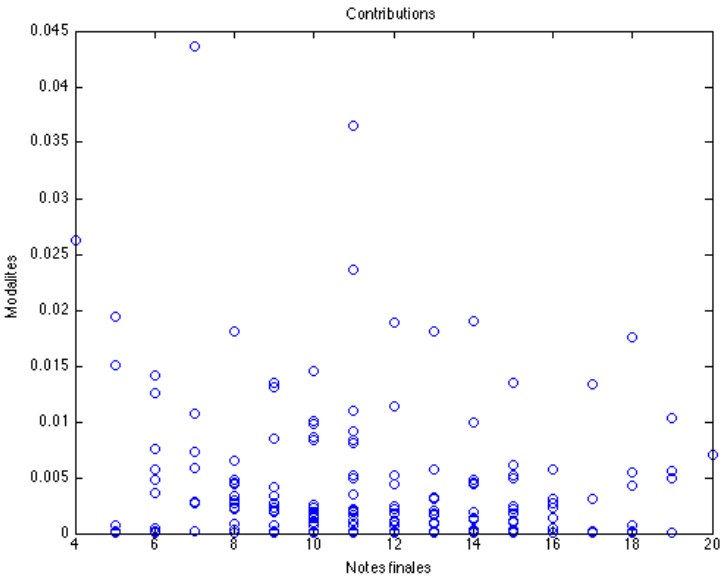

FIGURE 24 – Contributions selon les notes finales, sans les points aberrants

On obtient les paramètres suivants : $a_1 = 0,1117$; $a_2 = 0,8866$; $b = 0,1948$ ainsi qu'un coefficient de détermination très bon : $R^2 = 0,9347$. Une fois de plus, on s'aperçoit à l'aide de ces quelques graphiques que le modèle linéaire est très bien adapté à la situation et que l'utilisation des deux premières notes permet de prédire avec le plus de précision la note finales des élèves. La régression linéaire multiple est donc la mieux adaptée à notre problème et elle nous donne la formule suivante pour tenter de prédire la note finale des élèves si cette note n'est pas égale à 0 :

$$y = 0,1117 \cdot x_1 + 0,8866 \cdot x_2 + 0,1948$$

avec y la note finale, x_1 la première note et x_2 la seconde note.

2.4 Comparaison de boites à moustache

Lors de l'Analyse en Composantes Principales, nous avions conclu que les variables `Medu` et `Fedu` étaient corrélées positivement avec la note finale des élèves tandis que les variables `Failures` et `Age` étaient corrélées négativement avec cette dernière. Pour essayer de confirmer ces informations nous allons faire une comparaison de boites à moustache pour voir si les indicateurs (moyenne, médiane, et quartiles) diminuent ou augmentent en fonction des modalités des variables.

La variable `Failures`

Après avoir séparé les élèves en fonction de leur nombre de redoublements nous avons obtenu les boites à moustaches suivantes.

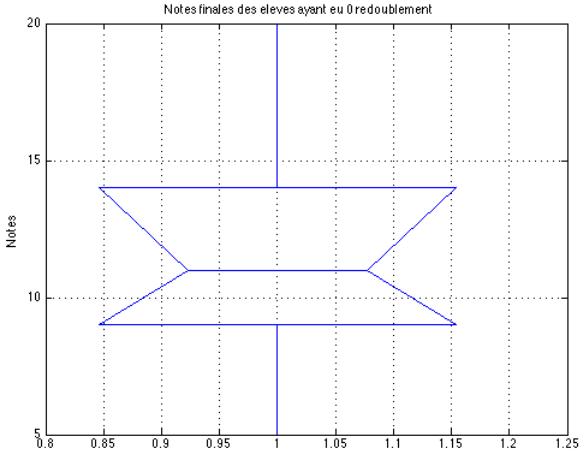

FIGURE 25 – Boîte à moustache des notes finales des élèves n'ayant jamais redoublé

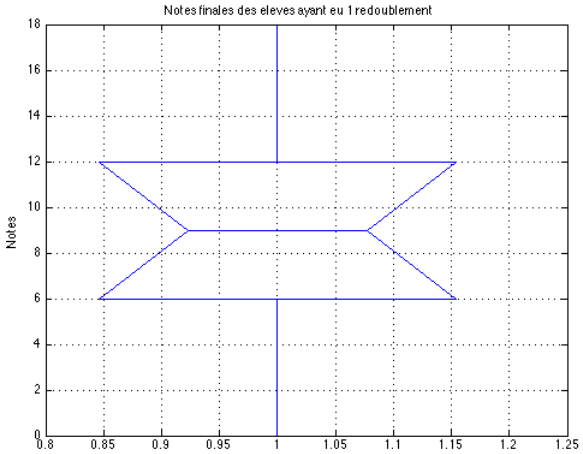

FIGURE 26 – Boîte à moustache des notes finales des élèves ayant redoublé une fois

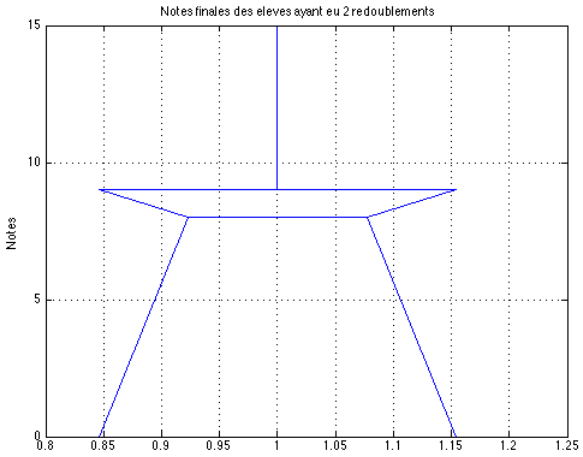

FIGURE 27 – Boîte à moustache des notes finales des élèves ayant redoublé deux fois

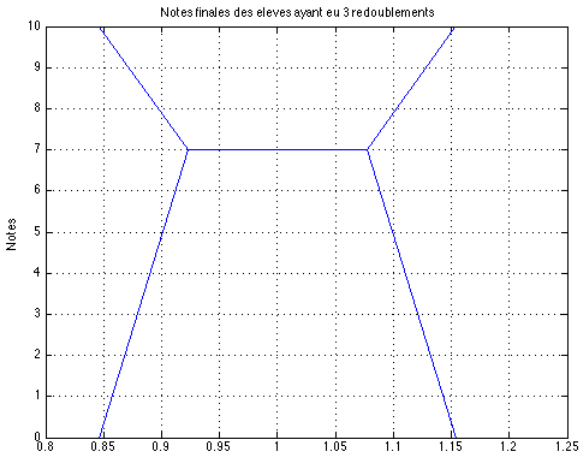

FIGURE 28 – Boîte à moustache des notes finales des élèves ayant redoublé trois fois

On constate que les principaux indicateurs (médiane et quartiles) ont tendance à diminuer plus le nombre de redoublement des élèves augmente. Ceci vient donc confirmer la thèse de la corrélation négative entre le nombre de redoublements des élèves et leurs notes finales. L'histogramme suivant concernant les moyennes des notes obtenues par chacun des groupes vient appuyer ce phénomène.

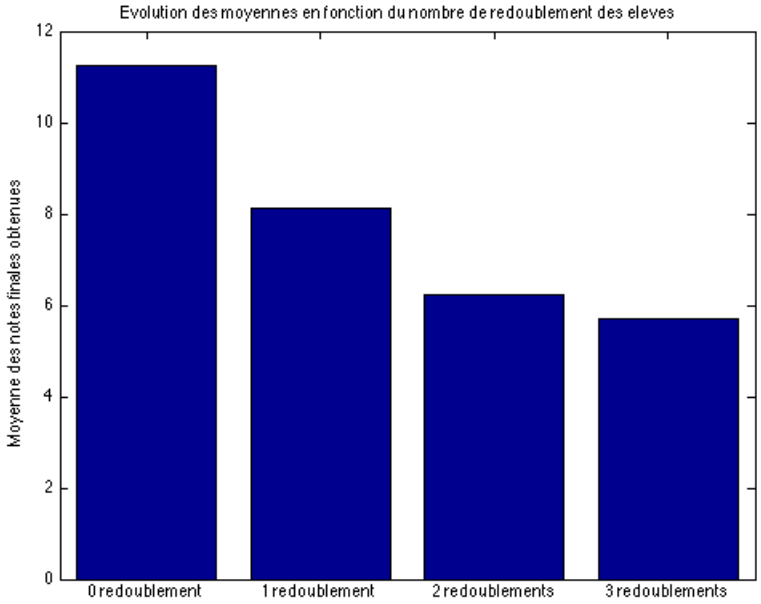

FIGURE 29 – Evolution des moyennes des élèves en fonction de leur nombre de redoublement

Nous pouvons donc affirmer que le nombre de redoublements a bien un effet négatif sur la note finale obtenue par les élèves. Plus le nombre de redoublements des élèves est élevé et plus leur note finale sera faible et inversement.

Les variables `Medu` et `Fedu`

Après avoir séparé les élèves en fonction du niveau d'éducation de leurs parents, on constate en comparant les boites à moustaches suivantes, que les médianes et quartiles des différents échantillons

tendent à augmenter à mesure que le niveau d'éducation des parents des élèves augmente. De même, l'évolution des moyennes reflète parfaitement ce phénomène et montre qu'il existe un lien évident entre les notes finales des élèves et le niveau d'éducation de leurs parents. Malheureusement, nous n'avons pas pu représenter le niveau d'éducation 0 étant donné que les élèves concernés étaient très peu nombreux. Cela ne nous permettait donc pas de les représenter au travers de leurs notes dans une boite à moustache.

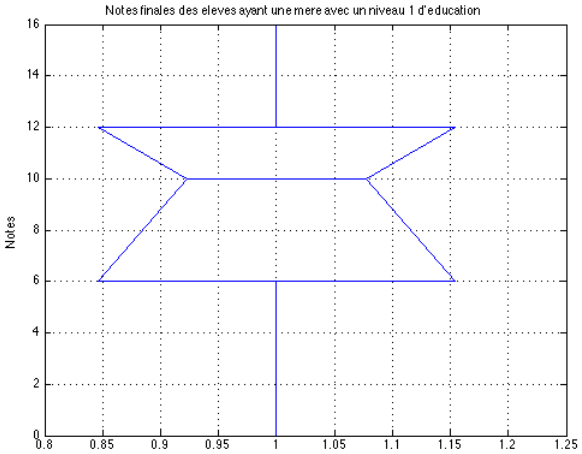

FIGURE 30 – Boîte à moustache des notes finales des élèves ayant une mère avec un niveau 1 d'éducation

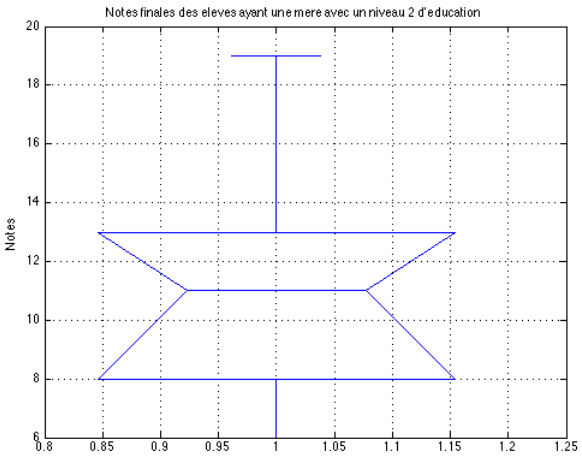

FIGURE 31 – Boîte à moustache des notes finales des élèves ayant une mère avec un niveau 2 d'éducation

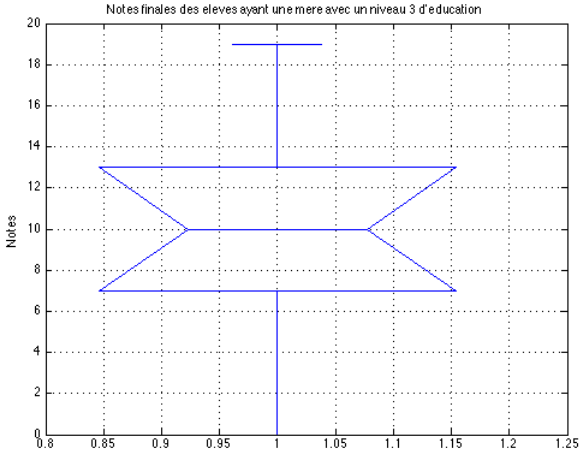

FIGURE 32 – Boîte à moustache des notes finales des élèves ayant une mère avec un niveau 3 d'éducation

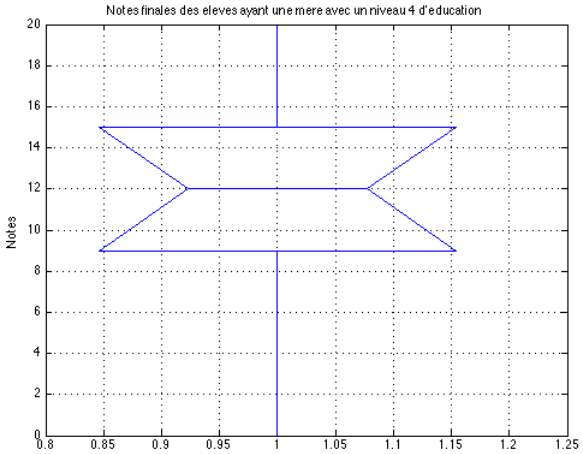

FIGURE 33 – Boîte à moustache des notes finales des élèves ayant une mère avec un niveau 4 d'éducation

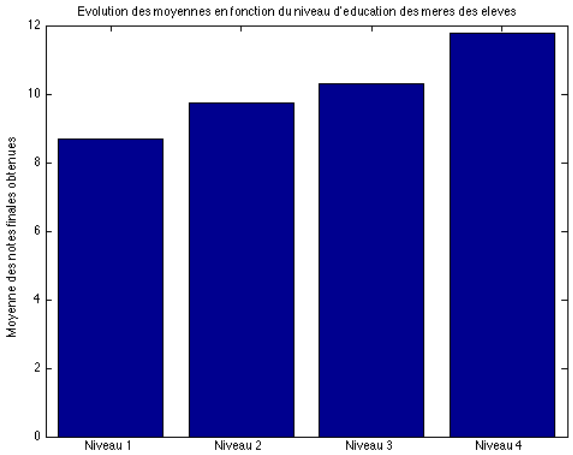

FIGURE 34 – Evolution des moyennes des élèves en fonction du niveau d'éducation des mères des élèves

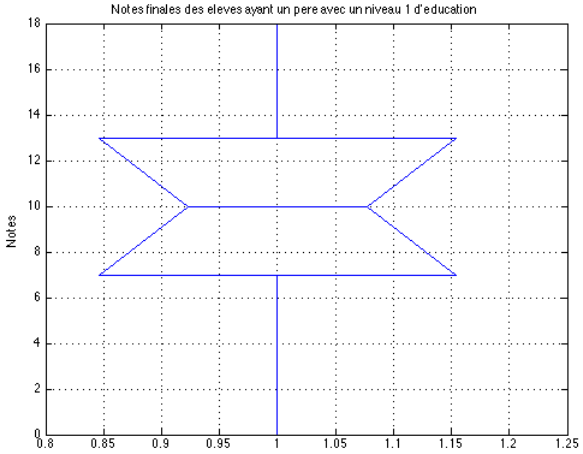

FIGURE 35 – Boîte à moustache des notes finales des élèves ayant un père avec un niveau 1 d'éducation

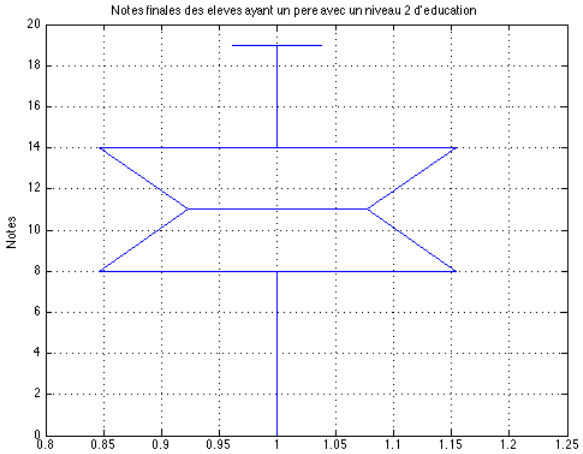

FIGURE 36 – Boîte à moustache des notes finales des élèves ayant un père avec un niveau 2 d'éducation

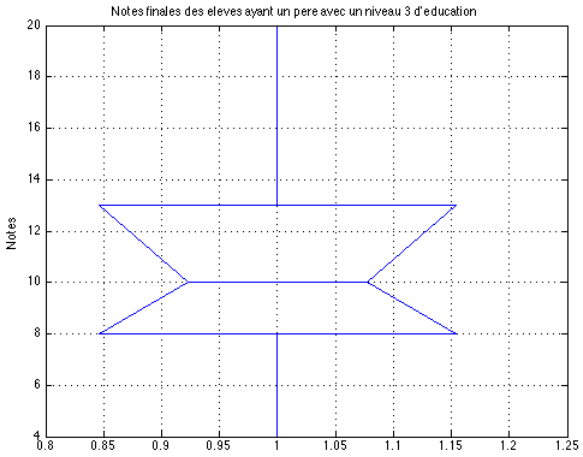

FIGURE 37 – Boîte à moustache des notes finales des élèves ayant un père avec un niveau 3 d'éducation

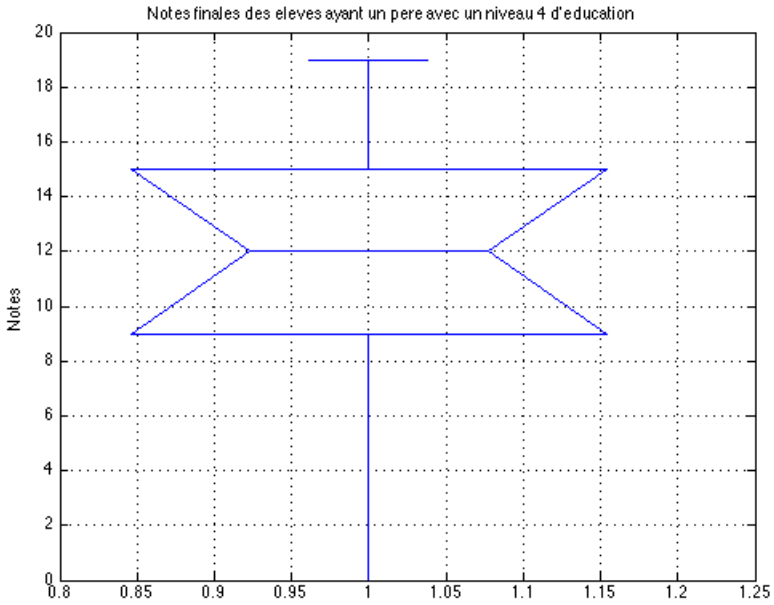

FIGURE 38 – Boîte à moustache des notes finales des élèves ayant un père avec un niveau 4 d'éducation

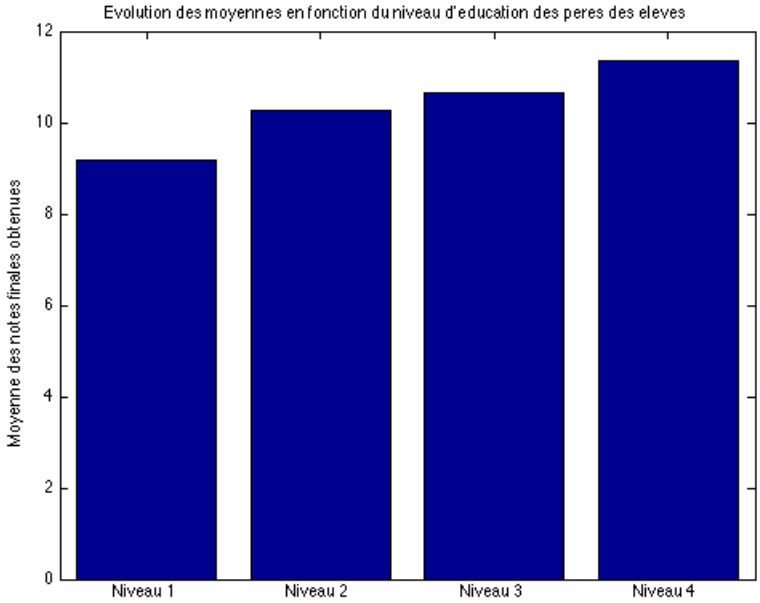

FIGURE 39 – Evolution des moyennes des élèves en fonction du niveau d'éducation des pères des élèves

Les variables `Medu` et `Fedu` sont donc bien corrélées positivement avec la variable contenant les notes finales des élèves. On est donc à présent en mesure d'affirmer que plus le niveau d'éducation des parents est élevé et plus la note finale de leur enfant sera elle aussi élevée, et inversement.

La variable `Age`

Nous avons dû une nouvelle fois séparer les élèves en plusieurs groupes distincts mais cette fois-ci en fonction de leurs âges. Cependant, les notes des élèves d'un âge strictement supérieur à 19 ans n'ont pu être représentées dans une boite à moustache car leur effectif était trop faible. Mais cela ne nous a en aucun cas empêché d'obtenir des résultats pour le moins satisfaisants. En effet, on remarque sur les graphiques qui vont suivre, que tous les indicateurs (moyenne médiane et quartiles) ont tendance à diminuer au fur et à mesure que l'âge des élèves augmente.

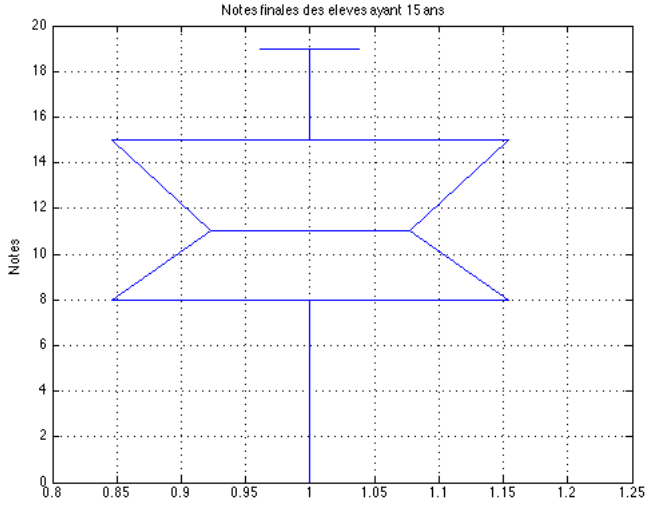

FIGURE 40 – Boîte à moustache des notes finales des élèves ayant 15 ans

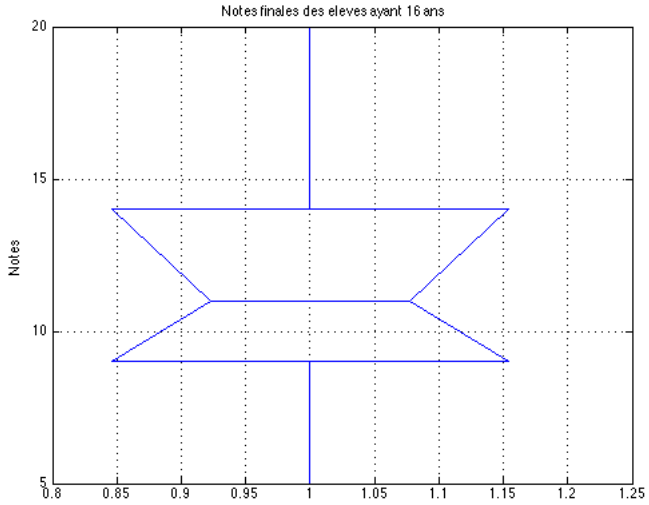

FIGURE 41 – Boîte à moustache des notes finales des élèves ayant 16 ans

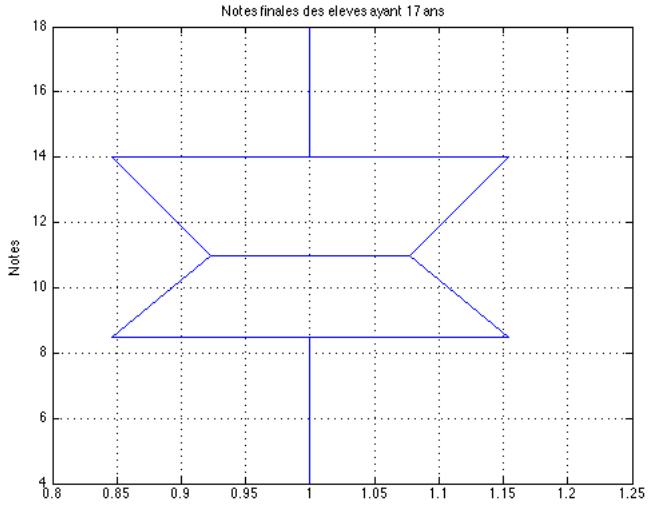

FIGURE 42 – Boîte à moustache des notes finales des élèves ayant 17 ans

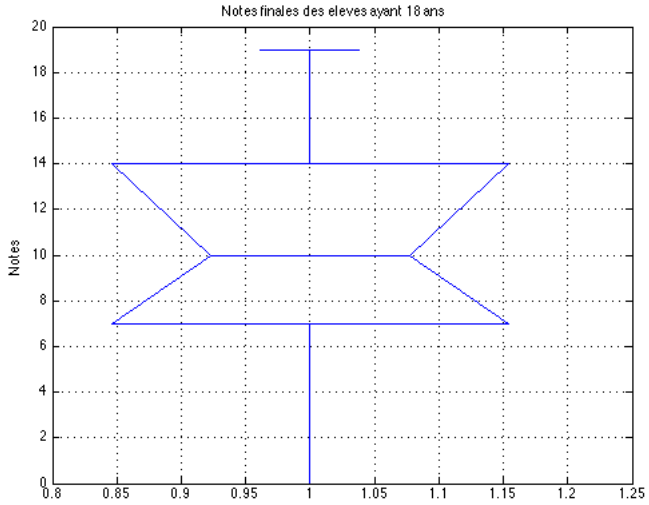

FIGURE 43 – Boîte à moustache des notes finales des élèves ayant 18 ans

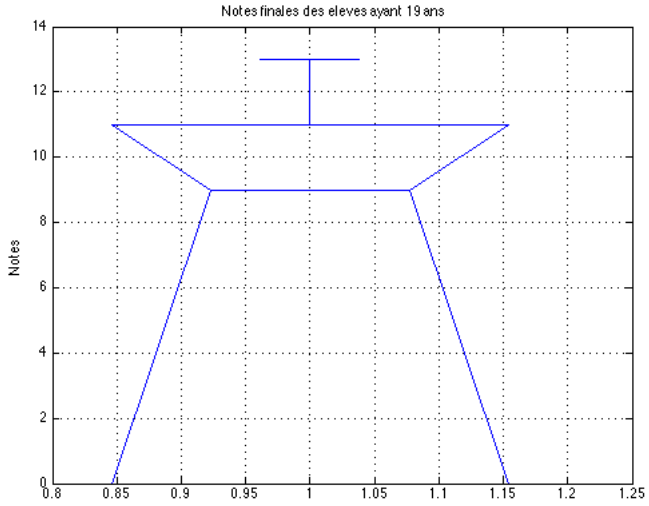

FIGURE 44 – Boîte à moustache des notes finales des élèves ayant 19 ans

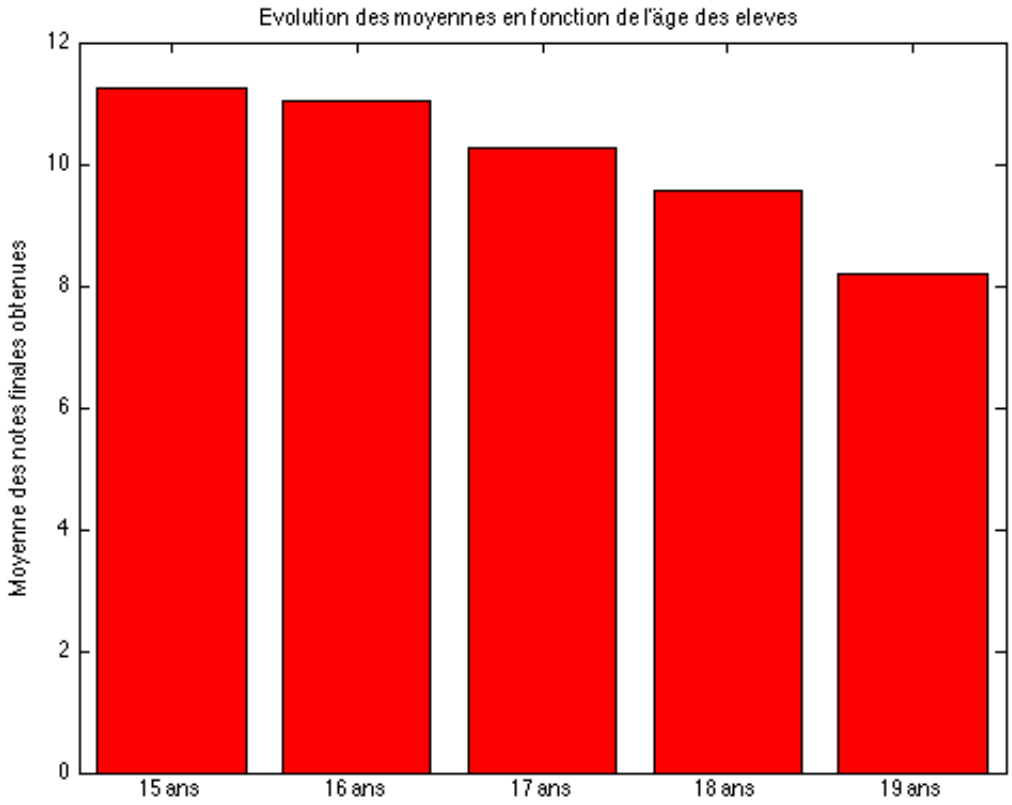

FIGURE 45 – Evolution des moyennes des élèves en fonction de leur âge

Nous sommes donc capables de dire que l'âge des élèves est corrélé négativement avec leur note finale. Ce qui signifie que plus l'âge des élèves est élevé et plus leur note finale sera faible et inversement. Ceci paraît logique car le nombre de redoublements a un rapport direct avec l'âge de l'élève. En effet, si on a déjà redoublé, on est plus vieux que ses camarades de classe, et de la même manière, si on est plus vieux, cela a de très fortes chances de signifier que l'on a déjà redoublé. De plus, comme le nombre de redoublements influence négativement la note finale alors il est très probable que l'âge en fasse de même.

Ce traitement de données nous a donc permis de tirer plusieurs enseignements. Tout abord nous avons vu que nous étions capables de prédire les notes finales des élèves (si celles-ci ne sont pas égales à 0) avec l'aide des deux premières notes obtenues par ces élèves. Par ailleurs, nous avons trouvé, avec les moyens qui sont les nôtres, que 4 variables avaient une influence sur la note finale obtenue par les élèves. En effet, nous avons découvert que `Medu` et `Fedu` étaient corrélées positivement avec la note finale, tandis que `Failures` et `Age` influençaient négativement cette dernière. Ces variables sont au nombre de 4 car nous avons fait le choix de ne tester que les variables ayant le plus de chances d'être corrélées (positivement ou négativement) avec la variable contenant les notes finales obtenues par les élèves. Ce qui signifie que des variables non traitées peuvent avoir elles aussi une influence sur la note finales des élèves.

3 Tests

3.1 Tests du χ^2

Nous réaliserons ici différents tests. Cependant, le raisonnement étant toujours analogue, nous n'allons détailler qu'une première fois les calculs.

3.1.1 Test du χ^2 détaillé sur les variables `Internet` et `Studytime`

Pour réaliser des tests du χ^2, il faut choisir deux variables qualitatives. Dans notre jeu de données, la plupart des variables sont qualitatives et il nous serait long et fastidieux de tester toutes les variables deux par deux. Ainsi, avec les avals de Messieurs DELPORTE, CANU et ROUSSELLE, nous avons décidé de ne tester que des couples de variables qui pouvait *a priori* être liées. Ainsi, comme premier test, nous avons décidé de tester les variables `Internet`, qui indique si un étudiant a accès à un Internet ou non, et `Studytime`, qui est une variable de catégories selon le temps que passe l'élève à étudier par semaine. Nous allons, une fois encore, développer les calculs réalisés à partir des données du fichier `student-mat.csv` mais le raisonnement est totalement analogue avec les données issues de l'autre fichier.

Etape 1 : construire le tableau de contingence

Tout d'abord, on fixe nos hypothèses de départ :

- « H_0 : le fait qu'un élève ait accès à Internet n'est pas en lien avec le temps qu'il passe à étudier chaque semaine ».
- « H_1 : le fait qu'un élève ait accès à Internet est lié avec le temps qu'il passe à étudier chaque semaine ».

Ensuite, on construit le tableau de contingence O des observations. On réalise pour cela plusieurs boucles, en passant par une matrice temporaire qui regroupe les deux variables à étudier, pour chercher les effectifs d'individus qui correspondent respectivement aux différents couples de variables.

Code 5 – Extrait des boucles permettant de construire le tableau de contingence du test du χ^2

```
15      temp=[TabModMat_Internet , Mat_Studytime];
16
17      O = zeros(2,4); % 2 lignes pour 'internet' et 4 colonnes pour 'studytime'
18
19      ind=[]; %Pour [0,1]
20      for i=1:length(temp)
21          if temp(i,:)==[0,1]
22              ind=[ind i];
23          end
24      end
25     O(1,1)=length(ind);
26
27      ind=[]; %Pour [0,2]
28      for i=1:length(temp)
29          if temp(i,:)==[0,2]
30              ind=[ind i];
31          end
32      end
33     O(1,2)=length(ind);
34
35      ind=[]; %Pour [0,3]
36      for i=1:length(temp)
37          if temp(i,:)==[0,3]
38              ind=[ind i];
39          end
40      end
```

```
41      O(1,3)=length(ind);
42
43      % Et ainsi de suite pour les autres couples
```

A la fin, on obtient le tableau de contingence (contenant des effectifs) suivant :

TABLE 6 – Tableau de contingence du test de χ^2 entre les variables `Internet` et `Studytime`

Studytime / *Internet*	**Moins de 2 heures**	**Entre 2 et 5 heures**	**Entre 5 et 10 heures**	**Plus de 10 heures**
Non	19	37	6	4
Oui	86	161	59	23

Etape 2 : on calcule les marginales

On calcule les marginales du tableau de contingence.

Code 6 – Calcul des marginales du tableau de contingence pour le test du χ^2

```
86      [I,J]=size(O); %[2,4]
87      nI=sum(O'); %profil ligne = [66 329]
88      nJ=sum(O); %profil colonne = [105 198 65 27]
89      n=sum(sum(O)); % = 395
```

Etape 3 : on calcule les $T_{i,j}$ pour créer le tableau des effectifs théoriques (en supposant l'indépendance)

On calcule les $T_{i,j} = n \cdot p_i \cdot p_j$ avec le code suivant :

Code 7 – Calcul des effectifs théoriques pour le test du χ^2

```
92      T=(nI'*nJ)/n; % Effectifs theoriques <==> Tij=P(i.)*P(.j)*n
93                                            %= Ni./n * N.j/n * n
94      % T/n donne les pourcentages
```

Etape 4 : on calcule la distance du χ^2

La distance du χ^2 entre les effectifs observés dans le tableau de contingence et les effectifs théoriques est défini par la formule :

$$D(O,T) = \sum_{i=1}^{I}\sum_{j=1}^{J} \frac{(O_{i,j} - T_{i,j})^2}{T_{i,j}}$$

On la calcule en *Matlab* par le code suivant :

Code 8 – Calcul de la distance du χ^2

```
97      D= sum(sum((O-T).^2./T)); %Distance du Chi2 : 3.3831
```

Etape 5 : calcul du degré de liberté du χ^2

Le degré de liberté du χ^2 est donné par la relation : $ddl = (I-1)\cdot(J-1)$. Ainsi, dans notre cas : `ddl=(I-1)*(J-1); %3`.

Etape 6 : On regarde dans les tables de la loi χ^2 à *ddl* degrés de liberté

Soit Z une variable aléatoire suivant la loi χ^2 à *ddl* degrés de liberté. On cherche la P-Valeur, définie ici par :

$$pval = P(Z \geq D(O,T))$$

Ainsi, pour notre cas, on obtient la P-Valeur par le code suivant :

Code 9 – Calcul de la P-Valeur avec les tables de la loi de $\chi 2$

```
103 pval=1-chi2cdf(3.3831,3) % pval =  0.33624
```

Etape 7 : conclure

On remarque que $pval = 0,33624 \geq 0,05$, on rejette donc H_0 et il y a lieu de remettre en cause l'indépendance des variables `Internet` et `Studytime`.

Après avoir réalisé le même test sur le second fichier avec ces deux variables, nous aboutissons à la même conclusion.

3.1.2 Test du χ^2 sur les variables `Romantic` et `Walc`

On réalise à présent un test du χ^2 sur les variables `Romantic`, qui indique si l'élève était dans une relation amoureuse au moment de l'enquête, et `Walc` qui indique la consommation hebdomadaire d'alcool de l'élève. On n'indiquera pas les calculs, ceux-ci étant similaires à ceux réalisés précédemment. En revanche, tous les calculs sont disponibles dans le fichier joint dans l'archive, intitulé `Test_Chi2.m`. Par ailleurs, les résultats qui vont être énoncés ici seront ceux issus des données du fichier `student-mat.csv`.

Le tableau de contingence obtenu est le suivant. Pour la variable `Walc` on rappelle que les modalités vont de « 1 » pour une consommation très faible à « 5 » pour une consommation importante d'alcool.

TABLE 7 – Tableau de contingence du test de χ^2 entre les variables `Romantic` et `Walc`

Romantic \ *Walc*	**1**	**2**	**3**	**4**	**5**
Non	100	56	52	38	17
Oui	51	29	28	13	11

On obtient alors la distance du χ^2 : $D(O,T) = 1.9912$, et on obtient que la P-Valeur vaut : $pval = 0.73738$. On déduit alors que, comme $pval \geq 0.05$, on rejette H_0 et on en déduit que les deux variables ne sont pas statistiquement indépendantes. Même constat après avoir testé ces mêmes variables dans le fichier `student-por.csv`.

3.1.3 Test du χ^2 sur les variables `Medu` et `Fedu`

On réalise maintenant un test du χ^2 sur les variables `Medu`, qui représente le niveau d'éducation de la mère, et `Fedu` qui représente le niveau d'éducation du père. On n'indiquera pas les calculs, ceux-ci étant similaires à ceux réalisés précédemment. En revanche, tous les calculs sont disponibles dans le fichier joint dans l'archive, intitulé `Test_Chi2.m`. Par ailleurs, les résultats qui vont être énoncés ici seront ceux issus des données du fichier `student-mat.csv`.

Le tableau de contingence obtenu est le suivant. Se référer à la page 5 pour la signification des différentes modalités.

On obtient alors la distance du χ^2 : $D(O,T) = 199.9773$, et on obtient que la P-Valeur vaut : $pval = 0$. On déduit alors que, comme $pval \leq 0.05$, On garde donc H_0 : il n'y a pas lieu de remettre en cause l'indépendance des variables `Medu` et `Fedu`. Même constat après avoir testé ces mêmes variables dans le fichier `student-por.csv`.

TABLE 8 – Tableau de contingence du test de χ^2 entre les variables `Medu` et `Fedu`

Medu \ *Fedu*	**0**	**1**	**2**	**3**	**4**
0	0	1	2	0	0
1	1	37	15	5	1
2	0	28	51	17	7
3	0	15	28	38	18
4	1	1	19	40	70

3.1.4 Test du χ^2 sur les variables `Mjob` et `Fjob`

On réalise ici un test du χ^2 sur les variables `Mjob`, qui représente le type d'emploi de la mère, et `Fjob` qui représente le type d'emploi du père. On n'indiquera pas les calculs, ceux-ci étant similaires à ceux réalisés précédemment. En revanche, tous les calculs sont disponibles dans le fichier joint dans l'archive, intitulé `Test_Chi2.m`. Par ailleurs, les résultats qui vont être énoncés ici seront ceux issus des données du fichier `student-mat.csv`.

Le tableau de contingence obtenu est le suivant. Se référer à la page 5 pour la signification des différentes modalités.

TABLE 9 – Tableau de contingence du test de χ^2 entre les variables `Mjob` et `Fjob`

Mjob \ *Fjob*	**0**	**1**	**2**	**3**	**4**
0	7	2	33	15	2
1	0	6	17	10	1
2	5	2	104	24	6
3	6	4	42	43	8
4	2	4	21	19	12

On obtient alors la distance du χ^2 : $D(O,T) = 73.3809$, et on obtient que la P-Valeur vaut : $pval = 2.5336 \cdot 10^{-9}$. On déduit alors que, comme $pval \leq 0.05$, On garde donc H_0 : il n'y a pas lieu de remettre en cause l'indépendance des variables `Mjob` et `Fjob`. Même constat après avoir testé ces mêmes variables dans le fichier `student-por.csv`.

3.1.5 Test du χ^2 sur les variables `Dalc` et `Walc`

On réalise ici un test du χ^2 sur les variables `Dalc`, qui représente la consommation quotidienne d'alcool de l'élève, et `Walc` qui représente sa consommation d'alcool hebdomadaire. On n'indiquera pas les calculs, ceux-ci étant similaires à ceux réalisés précédemment. En revanche, tous les calculs sont disponibles dans le fichier joint dans l'archive, intitulé `Test_Chi2.m`. Par ailleurs, les résultats qui vont être énoncés ici seront ceux issus des données du fichier `student-mat.csv`.

Le tableau de contingence obtenu est le suivant. Se référer à la page 7 pour la signification des différentes modalités.

TABLE 10 – Tableau de contingence du test de χ^2 entre les variables `Dalc` et `Walc`

Dalc \ *Walc*	**1**	**2**	**3**	**4**	**5**
1	150	65	42	15	4
2	1	18	29	22	5
3	0	1	8	11	6
4	0	1	1	3	4
5	0	0	0	0	9

On obtient alors la distance du χ^2 : $D(O,T) = 287.0019$, et on obtient que la P-Valeur vaut : $pval = 0$. On déduit alors que, comme $pval \leq 0.05$, On garde donc H_0 : il n'y a pas lieu de remettre en cause

l'indépendance des variables `Dalc` et `Walc`. Même constat après avoir testé ces mêmes variables dans le fichier `student-por.csv`.

Ici, le résultat paraît assez surprenant, car les variables semblent *a priori* pouvoir être liées. Pourtant, il semble bien au vu des résultats que la consommation quotidienne d'alcool et la consommation hebdomadaire ne soient pas liées.

3.2 Tests de Student

Pour réaliser des tests de STUDENT, il faut choisir deux variables quantitatives. Dans notre jeu de données, seules deux variables représentent réellement des quantités : `Age` et `Absences`. En effet, si nous transformons toutes les autres données (notamment celles qui sont d'origine en chaîne de caractères) en valeurs numériques, elles ne représentent pour autant pas des quantités, et un test de STUDENT n'a sur ces variables aucun intérêt. Par ailleurs, la plupart de nos variables représentent des catégories. En effet, à l'image de `Traveltime`, les valeurs que va prendre la variable seront « 1 » si le temps de voyage de l'élève est compris dans un intervalle de temps inférieur à 15 minutes, « 2 »si le temps est compris entre 15 et 30 minutes... La variable n'est pas quantitative dans le sens où elle ne représente pas réellement le temps de voyage : la variable n'est pas égale à 22 si l'élève met 22 minutes pour venir à l'école. Ainsi, il paraît assez peu intéressant de réaliser un test de STUDENT sur de telles variables.

Seules deux variables correspondent à des variables réellement quantitatives, l'âge de l'étudiant et son taux d'absentéisme, comme nous l'avons dit un peu plus tôt. Nous allons alors soumettre ces deux variables au test de STUDENT. Une fois de plus, nous ne représentons que le test de STUDENT réalisé sur les données du fichier `student-mat.csv`, mais les calculs sont identiques dans le fichier lié aux notes obtenues en cours de langue portugaise.

Etape 1 : formuler les hypothèses

Formulons les hypothèses dont nous allons essayer de trouver celle qui représente le plus la réalité :

- « H_0 : l'âge de l'élève n'est pas en lien avec son taux d'absentéisme ».
- « H_1 : l'âge de l'élève est lié à son taux d'absentéisme ».

Etape 2 : poser un modèle

Soit $\overline{x_{age}}$ (noté `Xb_age` dans le programme) la variable aléatoire représentant la moyenne des âges. On a alors :

$$\overline{x_{age}} \sim \mathcal{N}\left(\mu_{age}, \frac{\sigma^2_{age}}{n_{age}}\right)$$

On a alors `Xb_age=mean(Mat_Age); %16.6962 : age moyen`.

Par un raisonnement similaire, on a `Xb_absences=mean(Mat_Absences); %5.7089 : absenteisme moyen` .

On calcule alors le $\widehat{\sigma}^2$:

$$\widehat{\sigma}^2 = \frac{1}{n_{\text{age}} + n_{\text{absences}} - 2} \cdot \left(\sum_{i=1}^{n_{\text{age}}} (x_{\text{age},i} - \overline{x_{\text{age}}})^2 + \sum_{i=1}^{n_{\text{absences}}} (x_{\text{absences},i} - \overline{x_{\text{absences}}})^2 \right)$$

Code 10 – Calcul du $\widehat{\sigma}^2$ pour le test de STUDENT

```
28 %Calcul du Sigma2 :
29              sigma2=(1/(length(Mat_Absences) + length(Mat_Age)-2));
30              sigma2=sigma2*(sum((Mat_Absences-Xb_absences).^2)+sum((Mat_Age-Xb_age).^2));
31                                                          ↪%% 32.8389
```

On reformule alors les hypothèses :

- $H_0 : \mu_{\text{age}} = \mu_{\text{absences}}$.
- $H_1 : \mu_{\text{age}} \neq \mu_{\text{absences}}$.

Etape 3 : exhiber la statistique du test.

A ce moment, on peut calculer t :

$$t = \frac{\overline{x_{\text{age}}} - \overline{x_{\text{absences}}}}{\sqrt{\widehat{\sigma}^2 \left(\frac{1}{n_{\text{age}}} + \frac{1}{n_{\text{absences}}}\right)}}$$

et le nombre de degrés de liberté :

$$ddl = n_{\text{age}} + n_{\text{absences}} - 2$$

Code 11 – Calcul de t et du nombre de degrés de liberté pour le test de STUDENT

```
39 t=(Xb_age-Xb_absences) / (sqrt(sigma2*(1/length(Mat_Age) + 1/length(Mat_Absences))));
40                                          ↪ %26.9452
41 %Nombre de degres de liberte :
42 ddl = length(Mat_Age)+length(Mat_Absences)-2; %788
```

Etape 4 : calculer la P-Valeur

On calcule la p-valeur :

$$pval = 2 \cdot P(T \geq t)$$

Code 12 – Calcul de la P-Valeur pour le test de STUDENT

```
45 P_val=2*(1-cdf('t',26.9452, 788)) ; %0
```

Etape 5 : conclure

La P-Valeur étant inférieure à $\alpha = 0.05$, on garde donc l'hypothèse de départ H_0. Ainsi, on en déduit que l'âge des élèves n'est pas en lien avec leur taux d'absentéisme.

De prime abord, il paraissait presque évident que l'âge d'un élève n'influait pas sur son taux d'absentéisme. Le test de STUDENT nous l'a alors confirmé.

En réalisant les mêmes calculs avec le second fichier, on trouve $t = 69.4753$, $ddl = 1296$ et alors `P_val = 2*(1-cdf('t',69.4753,1296))%=0`. Pour ce second fichier, on en arrive à la même conclusion : l'âge n'influe pas sur l'absentéisme.

Conclusion

Ce projet nous a permis de mettre en œuvre quelques méthodes statistiques vues en cours tout en les adaptant à notre problème. Cela nous a donc appris à déceler quelles méthodes étaient les mieux adaptées à notre problématique de départ. En effet, le but de ce projet n'était pas selon nous de ressortir l'intégralité des méthodes vues en cours sans réfléchir. Nous avons souhaité rechercher quelles méthodes du cours convenaient le mieux à notre problème tout en ayant également accès à d'autres méthodes statistiques qui à nos yeux nous paraissaient intéressantes et utiles pour répondre à notre problématique.

Passons maintenant aux résultats plus concrets que nous a permis de révéler ce projet. Tout d'abord, nous avons appris grâce à quelques régressions linéaires qu'il était possible de prédire les notes finales des élèves à l'aide de leurs deux premières notes obtenues au cours de l'année. Dans un second temps, nous avons été en mesure, grâce à plusieurs méthodes (en l'occurrence ici, une analyse en composantes principales et des comparaisons de boîte à moustache), de montrer que certaines variables avaient une influence sur la note finale obtenue par les élèves. Nous avons notamment trouvé que les variables `Medu` et `Fedu` étaient corrélées positivement avec les notes finales des élèves tandis que les variables `Age` et `Failures` étaient corrélées négativement avec ces dernières. Dans un dernier temps nous avons décidé à l'aide de quelques tests de voir si certaines variables étaient liées entre elles. En effet, lors de la partie traitement nous annoncions qu'il paraissait logique que les variables `Age` et `Failures` soient liées. Cependant, nous avons prouvé à l'aide d'un test de STUDENT que ce n'était pas du tout le cas. Ainsi, ce que nous pensons être logique ne reflète pas toujours la réalité. De même, alors que nous pensions que les variables `Dalc` et `Walc` pouvaient avoir un lien entre elles et que les variables `Medu` et `Fedu` semblaient être corrélées selon notre ACP. Il s'est avéré que ce que nous pensions était faux.

En effet, nous avons été en mesure de montrer grâce au test du χ^2 que les couples de variables `Dalc`/`Walc`, `Medu`/`Fedu` ainsi que d'autres comme `Romantic`/`Walc` et `Mjob`/`Fjob` n'étaient pas liées.

Bien entendu, notre étude n'est en aucun cas exhaustive étant donné que notre base de données contient un très grand nombre de variables, il nous était donc impossible de traiter toutes les relations possibles entre chacune des variables. Nous avons préféré éclairer les relations entre les variables qui nous semblaient logiques ainsi que celles qui nous paraissaient intéressantes. Cependant, chacun peut s'il le souhaite approfondir notre sujet pour le traiter en intégralité, notamment en y apportant d'autres méthodes statistiques et en testant les relations entre les variables que nous n'avons pas pu traiter.

Liste des codes

Liste des tableaux

Table des figures

Annexes

Annexe A – `DonneesProjetM8.m`

```
%% Chargement des donnees de 'student-mat.csv'
dataMat=csvimport('student-mat.csv');

% Donnees numeriques
Mat_Age=cell2mat(dataMat(2:end,3));
Mat_Medu=cell2mat(dataMat(2:end,7));
Mat_Fedu=cell2mat(dataMat(2:end,8));
Mat_Traveltime=cell2mat(dataMat(2:end,13));
Mat_Studytime=cell2mat(dataMat(2:end,14));
Mat_Failures=cell2mat(dataMat(2:end,15));
Mat_Famrel=cell2mat(dataMat(2:end,24));
Mat_Freetime=cell2mat(dataMat(2:end,25));
Mat_Goout=cell2mat(dataMat(2:end,26));
Mat_Dalc=cell2mat(dataMat(2:end,27));
Mat_Walc=cell2mat(dataMat(2:end,28));
Mat_Health=cell2mat(dataMat(2:end,29));
Mat_Absences=cell2mat(dataMat(2:end,30));
Mat_G1=cell2mat(dataMat(2:end,31));
Mat_G2=cell2mat(dataMat(2:end,32));
Mat_G3=cell2mat(dataMat(2:end,33));

% Donnees en chaines de caracteres
Mat_School=char(dataMat(2:end,1));
Mat_Sex=char(dataMat(2:end,2));
Mat_Address=char(dataMat(2:end,4));
Mat_Famsize=char(dataMat(2:end,5));
Mat_Pstatus=char(dataMat(2:end,6));
Mat_Mjob=char(dataMat(2:end,9));
Mat_Fjob=char(dataMat(2:end,10));
Mat_Reason=char(dataMat(2:end,11));
Mat_Guardian=char(dataMat(2:end,12));
Mat_Schoolsup=char(dataMat(2:end,16));
Mat_Famsup=char(dataMat(2:end,17));
Mat_Paid=char(dataMat(2:end,18));
Mat_Activities=char(dataMat(2:end,19));
Mat_Nursery=char(dataMat(2:end,20));
Mat_Higher=char(dataMat(2:end,21));
Mat_Internet=char(dataMat(2:end,22));
Mat_Romantic=char(dataMat(2:end,23));



%% Chargement des donnees de 'student-por.csv'
dataPor=csvimport('student-por.csv');

% Donnees numeriques
Por_Age=cell2mat(dataPor(2:end,3));
Por_Medu=cell2mat(dataPor(2:end,7));
Por_Fedu=cell2mat(dataPor(2:end,8));
Por_Traveltime=cell2mat(dataPor(2:end,13));
Por_Studytime=cell2mat(dataPor(2:end,14));
Por_Failures=cell2mat(dataPor(2:end,15));
Por_Famrel=cell2mat(dataPor(2:end,24));
Por_Freetime=cell2mat(dataPor(2:end,25));
Por_Goout=cell2mat(dataPor(2:end,26));
Por_Dalc=cell2mat(dataPor(2:end,27));
```

```
Por_Walc=cell2mat(dataPor(2:end,28));
Por_Health=cell2mat(dataPor(2:end,29));
Por_Absences=cell2mat(dataPor(2:end,30));
Por_G1=cell2mat(dataPor(2:end,31));
Por_G2=cell2mat(dataPor(2:end,32));
Por_G3=cell2mat(dataPor(2:end,33));

% Donnees en chaines de caracteres
Por_School=char(dataPor(2:end,1));
Por_Sex=char(dataPor(2:end,2));
Por_Address=char(dataPor(2:end,4));
Por_Famsize=char(dataPor(2:end,5));
Por_Pstatus=char(dataPor(2:end,6));
Por_Mjob=char(dataPor(2:end,9));
Por_Fjob=char(dataPor(2:end,10));
Por_Reason=char(dataPor(2:end,11));
Por_Guardian=char(dataPor(2:end,12));
Por_Schoolsup=char(dataPor(2:end,16));
Por_Famsup=char(dataPor(2:end,17));
Por_Paid=char(dataPor(2:end,18));
Por_Activities=char(dataPor(2:end,19));
Por_Nursery=char(dataPor(2:end,20));
Por_Higher=char(dataPor(2:end,21));
Por_Internet=char(dataPor(2:end,22));
Por_Romantic=char(dataPor(2:end,23));


clear dataMat
clear dataPor

%% Rangement modalites (chaines) fichier 'student_mat.csv'

%Rangement modalites Mat_School (0=MS / 1=GP)
TabModMat_School=ones(length(Mat_School),1);
for i=1:length(Mat_School)
    ind=[];
    if (Mat_School(i,:)=='MS')
       ind=[ind ;i];
       TabModMat_School(ind,1)=0;
    end
end

%Rangement modalites Mat_Sex (0=F / 1=M)
TabModMat_Sex=ones(length(Mat_Sex),1);
for i=1:length(Mat_Sex)
    ind=[];
    if (Mat_Sex(i,:)=='F')
       ind=[ind ;i];
       TabModMat_Sex(ind,1)=0;
    end
end

%Rangement modalites Mat_Address (0=U / 1=R)
TabModMat_Address=ones(length(Mat_Address),1);
for i=1:length(Mat_Address)
    ind=[];
    if (Mat_Address(i,:)=='U')
       ind=[ind ;i];
       TabModMat_Address(ind,1)=0;
    end
end
```

```
118
119 %Rangement modalites Mat_Famsize (0=LE3 / 1=GT3)
120 TabModMat_Famsize=ones(length(Mat_Famsize),1);
121 for i=1:length(Mat_Famsize)
122     ind=[];
123     if (Mat_Famsize(i,:)=='LE3')
124        ind=[ind ;i];
125        TabModMat_Famsize(ind,1)=0;
126     end
127 end
128
129 %Rangement modalites Mat_Pstatus (0=A / 1=T)
130 TabModMat_Pstatus=ones(length(Mat_Pstatus),1);
131 for i=1:length(Mat_Pstatus)
132     ind=[];
133     if (Mat_Pstatus(i,:)=='A')
134        ind=[ind ;i];
135        TabModMat_Pstatus(ind,1)=0;
136     end
137 end
138
139 %Rangement modalites Mat_Mjob (0=at_home / 1=health / 2=other / 3=services / 4=teacher)
140 TabModMat_Mjob=ones(length(Mat_Mjob),1);
141 for i=1:length(Mat_Mjob)
142     ind=[];
143     if (Mat_Mjob(i,:)=='at_home ')
144        ind=[ind ;i];
145        TabModMat_Mjob(ind,1)=0;
146     end
147     if (Mat_Mjob(i,:)=='other   ')
148        ind=[ind ;i];
149        TabModMat_Mjob(ind,1)=2;
150     end
151     if (Mat_Mjob(i,:)=='services')
152        ind=[ind ;i];
153        TabModMat_Mjob(ind,1)=3;
154     end
155     if (Mat_Mjob(i,:)=='teacher ')
156        ind=[ind ;i];
157        TabModMat_Mjob(ind,1)=4;
158     end
159 end
160
161 %Rangement modalites Mat_Fjob (0=at_home / 1=health / 2=other / 3=services / 4=teacher)
162 TabModMat_Fjob=ones(length(Mat_Fjob),1);
163 for i=1:length(Mat_Fjob)
164     ind=[];
165     if (Mat_Fjob(i,:)=='at_home ')
166        ind=[ind ;i];
167        TabModMat_Fjob(ind,1)=0;
168     end
169     if (Mat_Fjob(i,:)=='other   ')
170        ind=[ind ;i];
171        TabModMat_Fjob(ind,1)=2;
172     end
173     if (Mat_Fjob(i,:)=='services')
174        ind=[ind ;i];
175        TabModMat_Fjob(ind,1)=3;
176     end
177     if (Mat_Fjob(i,:)=='teacher ')
178        ind=[ind ;i];
```

```
179         TabModMat_Fjob(ind,1)=4;
180     end
181 end
182
183 %Rangement modalites Mat_Reason (0=course / 1=home / 2=other / 3=reputation)
184 TabModMat_Reason=ones(length(Mat_Reason),1);
185 for i=1:length(Mat_Reason)
186     ind=[];
187     if (Mat_Reason(i,:)=='course    ')
188         ind=[ind ;i];
189         TabModMat_Reason(ind,1)=0;
190     end
191     if (Mat_Reason(i,:)=='other     ')
192         ind=[ind ;i];
193         TabModMat_Reason(ind,1)=2;
194     end
195     if (Mat_Reason(i,:)=='reputation')
196         ind=[ind ;i];
197         TabModMat_Reason(ind,1)=3;
198     end
199 end
200
201 %Rangement modalites Mat_Guardian (0=father / 1=mother / 2=other)
202 TabModMat_Guardian=ones(length(Mat_Guardian),1);
203 for i=1:length(Mat_Guardian)
204     ind=[];
205     if (Mat_Guardian(i,:)=='father')
206         ind=[ind ;i];
207         TabModMat_Guardian(ind,1)=0;
208     end
209     if (Mat_Guardian(i,:)=='other ')
210         ind=[ind ;i];
211         TabModMat_Guardian(ind,1)=2;
212     end
213 end
214
215 %Rangement modalites Mat_Schoolsup (0=no / 1=yes)
216 TabModMat_Schoolsup=ones(length(Mat_Schoolsup),1);
217 for i=1:length(Mat_Schoolsup)
218     ind=[];
219     if (Mat_Schoolsup(i,:)=='no ')
220         ind=[ind ;i];
221         TabModMat_Schoolsup(ind,1)=0;
222     end
223 end
224
225 %Rangement modalites Mat_Famsup (0=no / 1=yes)
226 TabModMat_Famsup=ones(length(Mat_Famsup),1);
227 for i=1:length(Mat_Famsup)
228     ind=[];
229     if (Mat_Famsup(i,:)=='no ')
230         ind=[ind ;i];
231         TabModMat_Famsup(ind,1)=0;
232     end
233 end
234
235 %Rangement modalites Mat_Paid (0=no / 1=yes)
236 TabModMat_Paid=ones(length(Mat_Paid),1);
237 for i=1:length(Mat_Paid)
238     ind=[];
239     if (Mat_Paid(i,:)=='no ')
```

```
240         ind=[ind ;i];
241         TabModMat_Paid(ind,1)=0;
242     end
243 end
244
245 %Rangement modalites Mat_Activities (0=no / 1=yes)
246 TabModMat_Activities=ones(length(Mat_Activities),1);
247 for i=1:length(Mat_Activities)
248     ind=[];
249     if (Mat_Activities(i,:)=='no ')
250         ind=[ind ;i];
251         TabModMat_Activities(ind,1)=0;
252     end
253 end
254
255 %Rangement modalites Mat_Nursery (0=no / 1=yes)
256 TabModMat_Nursery=ones(length(Mat_Nursery),1);
257 for i=1:length(Mat_Nursery)
258     ind=[];
259     if (Mat_Nursery(i,:)=='no ')
260         ind=[ind ;i];
261         TabModMat_Nursery(ind,1)=0;
262     end
263 end
264
265 %Rangement modalites Mat_Higher (0=no / 1=yes)
266 TabModMat_Higher=ones(length(Mat_Higher),1);
267 for i=1:length(Mat_Higher)
268     ind=[];
269     if (Mat_Higher(i,:)=='no ')
270         ind=[ind ;i];
271         TabModMat_Higher(ind,1)=0;
272     end
273 end
274
275 %Rangement modalites Mat_Internet (0=no / 1=yes)
276 TabModMat_Internet=ones(length(Mat_Internet),1);
277 for i=1:length(Mat_Internet)
278     ind=[];
279     if (Mat_Internet(i,:)=='no ')
280         ind=[ind ;i];
281         TabModMat_Internet(ind,1)=0;
282     end
283 end
284
285 %Rangement modalites Mat_Romantic (0=no / 1=yes)
286 TabModMat_Romantic=ones(length(Mat_Romantic),1);
287 for i=1:length(Mat_Romantic)
288     ind=[];
289     if (Mat_Romantic(i,:)=='no ')
290         ind=[ind ;i];
291         TabModMat_Romantic(ind,1)=0;
292     end
293 end
294
295
296 %% Rangement modalites (chaines) fichier 'student_por.csv'
297
298 %Rangement modalites Por_School (0=MS / 1=GP)
299 TabModPor_School=ones(length(Por_School),1);
300 for i=1:length(Por_School)
```

```
301      ind=[];
302      if (Por_School(i,:)=='MS')
303         ind=[ind ;i];
304         TabModPor_School(ind,1)=0;
305      end
306  end
307
308  %Rangement modalites Por_Sex (0=F / 1=M)
309  TabModPor_Sex=ones(length(Por_Sex),1);
310  for i=1:length(Por_Sex)
311      ind=[];
312      if (Por_Sex(i,:)=='F')
313         ind=[ind ;i];
314         TabModPor_Sex(ind,1)=0;
315      end
316  end
317
318  %Rangement modalites Por_Address (0=U / 1=R)
319  TabModPor_Address=ones(length(Por_Address),1);
320  for i=1:length(Por_Address)
321      ind=[];
322      if (Por_Address(i,:)=='U')
323         ind=[ind ;i];
324         TabModPor_Address(ind,1)=0;
325      end
326  end
327
328  %Rangement modalites Por_Famsize (0=LE3 / 1=GT3)
329  TabModPor_Famsize=ones(length(Por_Famsize),1);
330  for i=1:length(Por_Famsize)
331      ind=[];
332      if (Por_Famsize(i,:)=='LE3')
333         ind=[ind ;i];
334         TabModPor_Famsize(ind,1)=0;
335      end
336  end
337
338  %Rangement modalites Por_Pstatus (0=A / 1=T)
339  TabModPor_Pstatus=ones(length(Por_Pstatus),1);
340  for i=1:length(Por_Pstatus)
341      ind=[];
342      if (Por_Pstatus(i,:)=='A')
343         ind=[ind ;i];
344         TabModPor_Pstatus(ind,1)=0;
345      end
346  end
347
348  %Rangement modalites Por_Mjob (0=at_home / 1=health / 2=other / 3=services / 4=teacher)
349  TabModPor_Mjob=ones(length(Por_Mjob),1);
350  for i=1:length(Por_Mjob)
351      ind=[];
352      if (Por_Mjob(i,:)=='at_home ')
353         ind=[ind ;i];
354         TabModPor_Mjob(ind,1)=0;
355      end
356      if (Por_Mjob(i,:)=='other   ')
357         ind=[ind ;i];
358         TabModPor_Mjob(ind,1)=2;
359      end
360      if (Por_Mjob(i,:)=='services')
361         ind=[ind ;i];
```

```
362         TabModPor_Mjob(ind,1)=3;
363     end
364     if (Por_Mjob(i,:)=='teacher ')
365         ind=[ind ;i];
366         TabModPor_Mjob(ind,1)=4;
367     end
368 end
369
370 %Rangement modalites Por_Fjob (0=at_home / 1=health / 2=other / 3=services / 4=teacher)
371
372 TabModPor_Fjob=ones(length(Por_Fjob),1);
373 for i=1:length(Por_Fjob)
374     ind=[];
375     if (Por_Fjob(i,:)=='at_home ')
376
377         ind=[ind ;i];
378         TabModPor_Fjob(ind,1)=0;
379     end
380     if (Por_Fjob(i,:)=='other   ')
381         ind=[ind ;i];
382         TabModPor_Fjob(ind,1)=2;
383     end
384     if (Por_Fjob(i,:)=='services')
385         ind=[ind ;i];
386         TabModPor_Fjob(ind,1)=3;
387     end
388     if (Por_Fjob(i,:)=='teacher ')
389         ind=[ind ;i];
390         TabModPor_Fjob(ind,1)=4;
391     end
392 end
393
394 %Rangement modalites Por_Reason (0=course / 1=home / 2=other / 3=reputation)
395 TabModPor_Reason=ones(length(Por_Reason),1);
396 for i=1:length(Por_Reason)
397     ind=[];
398     if (Por_Reason(i,:)=='course    ')
399         ind=[ind ;i];
400         TabModPor_Reason(ind,1)=0;
401     end
402     if (Por_Reason(i,:)=='other     ')
403         ind=[ind ;i];
404         TabModPor_Reason(ind,1)=2;
405     end
406     if (Por_Reason(i,:)=='reputation')
407         ind=[ind ;i];
408         TabModPor_Reason(ind,1)=3;
409     end
410 end
411
412 %Rangement modalites Por_Guardian (0=father / 1=mother / 2=other)
413 TabModPor_Guardian=ones(length(Por_Guardian),1);
414 for i=1:length(Por_Guardian)
415     ind=[];
416     if (Por_Guardian(i,:)=='father')
417         ind=[ind ;i];
418         TabModPor_Guardian(ind,1)=0;
419     end
420     if (Por_Guardian(i,:)=='other ')
421         ind=[ind ;i];
422         TabModPor_Guardian(ind,1)=2;
```

```
423     end
424 end
425 
426 %Rangement modalites Por_Schoolsup (0=no / 1=yes)
427 TabModPor_Schoolsup=ones(length(Por_Schoolsup),1);
428 for i=1:length(Por_Schoolsup)
429     ind=[];
430     if (Por_Schoolsup(i,:)=='no ')
431        ind=[ind ;i];
432        TabModPor_Schoolsup(ind,1)=0;
433     end
434 end
435 
436 %Rangement modalites Por_Famsup (0=no / 1=yes)
437 TabModPor_Famsup=ones(length(Por_Famsup),1);
438 for i=1:length(Por_Famsup)
439     ind=[];
440     if (Por_Famsup(i,:)=='no ')
441        ind=[ind ;i];
442        TabModPor_Famsup(ind,1)=0;
443     end
444 end
445 
446 %Rangement modalites Por_Paid (0=no / 1=yes)
447 TabModPor_Paid=ones(length(Por_Paid),1);
448 for i=1:length(Por_Paid)
449     ind=[];
450     if (Por_Paid(i,:)=='no ')
451        ind=[ind ;i];
452        TabModPor_Paid(ind,1)=0;
453     end
454 end
455 
456 %Rangement modalites Por_Activities (0=no / 1=yes)
457 TabModPor_Activities=ones(length(Por_Activities),1);
458 for i=1:length(Por_Activities)
459     ind=[];
460     if (Por_Activities(i,:)=='no ')
461        ind=[ind ;i];
462        TabModPor_Activities(ind,1)=0;
463     end
464 end
465 
466 %Rangement modalites Por_Nursery (0=no / 1=yes)
467 TabModPor_Nursery=ones(length(Por_Nursery),1);
468 for i=1:length(Por_Nursery)
469     ind=[];
470     if (Por_Nursery(i,:)=='no ')
471        ind=[ind ;i];
472        TabModPor_Nursery(ind,1)=0;
473     end
474 end
475 
476 %Rangement modalites Por_Higher (0=no / 1=yes)
477 TabModPor_Higher=ones(length(Por_Higher),1);
478 for i=1:length(Por_Higher)
479     ind=[];
480     if (Por_Higher(i,:)=='no ')
481        ind=[ind ;i];
482        TabModPor_Higher(ind,1)=0;
483     end
```

```
484 end
485
486 %Rangement modalites Por_Internet (0=no / 1=yes)
487 TabModPor_Internet=ones(length(Por_Internet),1);
488 for i=1:length(Por_Internet)
489     ind=[];
490     if (Por_Internet(i,:)=='no ')
491        ind=[ind ;i];
492        TabModPor_Internet(ind,1)=0;
493     end
494 end
495
496 %Rangement modalites Por_Romantic (0=no / 1=yes)
497
498 TabModPor_Romantic=ones(length(Por_Romantic),1);
499 for i=1:length(Por_Romantic)
500     ind=[];
501     if (Por_Romantic(i,:)=='no ')
502        ind=[ind ;i];
503        TabModPor_Romantic(ind,1)=0;
504     end
505 end
506
507
508
509
510 clear ind
```

Annexe B – `Traitement.m`

```
run('DonneesProjetM8.m');
X_Mat=[TabModMat_School,TabModMat_Sex,Mat_Age,TabModMat_Address,TabModMat_Famsize,
    TabModMat_Pstatus,Mat_Medu,Mat_Fedu,TabModMat_Mjob,TabModMat_Fjob,TabModMat_Reason,
    TabModMat_Guardian,Mat_Traveltime,Mat_Studytime,Mat_Failures,TabModMat_Schoolsup,
    TabModMat_Famsup,TabModMat_Paid,TabModMat_Activities,TabModMat_Nursery,
    TabModMat_Higher,TabModMat_Internet,TabModMat_Romantic,Mat_Famrel,Mat_Freetime,
    Mat_Goout,Mat_Dalc,Mat_Walc,Mat_Health,Mat_Absences,Mat_G1,Mat_G2,Mat_G3];
Var_Name=[' School    ';' Sex       ';' Age       ';' Address   ';' Famsize   ';
' Pstatus   ';' Medu      ';' Fedu      ';' Mjob      ';' Fjob      ';' Reason    ';
' Guardian  ';' Traveltime';' Studytime ';' Failures  ';' Schoolsup ';' Famsup    ';
' Paid      ';' Activities';' Nursery   ';' Higher    ';' Internet  ';' Romantic  ';
' Famrel    ';' Freetime  ';' Goout     ';' Dalc      ';' Walc      ';' Health    ';
' Absences  ';' 1ereNote  ';' 2emeNote  ';' NoteFinale'];
[n p]=size(X_Mat);

%% Etude des notes finales
Mat_G3_classe=sort(Mat_G3);
[modalites N]=unique(Mat_G3_classe);
effectifs=[diff(N);1];
effectifs_cumules=cumsum(effectifs);
frequences=effectifs/n;
frequences_cumulees=effectifs_cumules/n;
frequences_corrigees(1,1)=frequences_cumulees(1);
for i=2:length(modalites)
    frequences_corrigees(i,1)=frequences_cumulees(i)-0.5*(frequences_cumulees(i)-
        frequences_cumulees(i-1));
end;
figure;
%plot(modalites,frequences_cumulees);
set(gcf,'Color',[1,1,1])
plot(modalites,frequences_cumulees,'o-') ; hold on
for i=1:length(modalites)-1
    plot([modalites(i) modalites(i+1)],[frequences_cumulees(i) frequences_cumulees(i)],'r
        ');
    plot([modalites(i+1) modalites(i+1)],[frequences_cumulees(i) frequences_cumulees(i+1)
        ],'r');
end
title('Fonction de repartition des notes finales des eleves');
xlabel('Notes finales');
ylabel('Frequences cumulees');
grid on;
moyenne_notes=mean(Mat_G3);
figure;
boxPlot(Mat_G3);
title('Notes finales des eleves');
ylabel('Notes');
grid on;
indice_echec=max(find(modalites<10));
taux_echec=frequences_cumulees(indice_echec)*100;
effectif_echec=effectifs_cumules(indice_echec);
histogramme=[effectif_echec,n-effectif_echec];
figure;
bar(histogramme,'r');
title('Validation de la matiere');
ylabel('Nombre d''eleves');
set(gca,'XTickLabel',{'echec','reussite'})
indice_passable_max=max(find(modalites>=10 & modalites<12));
effectif_passable=effectifs_cumules(indice_passable_max)-effectifs_cumules(indice_echec);
```

```
indice_assezbien_max=max(find(modalites>=12 & modalites<14));
effectif_assezbien=effectifs_cumules(indice_assezbien_max)-effectifs_cumules(
    indice_passable_max);
indice_bien_max=max(find(modalites>=14 & modalites<16));
effectif_bien=effectifs_cumules(indice_bien_max)-effectifs_cumules(indice_assezbien_max);
indice_tresbien_max=max(find(modalites>=16 & modalites<18));
effectif_tresbien=effectifs_cumules(indice_tresbien_max)-effectifs_cumules(
    indice_bien_max);
effectif_excellent=n-effectifs_cumules(indice_tresbien_max);
histogramme=[effectif_echec,effectif_passable,effectif_assezbien,effectif_bien,
    effectif_tresbien,effectif_excellent];
figure;
bar(histogramme);
title('Mentions obtenues par les eleves');
ylabel('Nombre d''eleves');
set(gca,'XTickLabel',{'echec','aucune','assez bien','bien','tres bien','excellent'});
%% Analyse en composantes principales
moyenne=ones(n,1)*mean(X_Mat);
ecart_type=ones(n,1)*std(X_Mat);
Xc_Mat=(X_Mat-moyenne);% matrice centree
Xn_Mat=Xc_Mat./ecart_type;% matrice centree et reduite
rho=Xn_Mat'*Xn_Mat/n;% matrice des correlations
[V D]=eig(Xn_Mat'*Xn_Mat);% calcul des vecteurs et valeurs propres
lambda=diag(D);% valeurs propres
info=lambda/sum(lambda)*100;% pourcentage d'information porte par chacune des composantes
    principales
Vn = V*sqrt(D(1:p,1:p)/n);
t=0:0.01:2*pi+0,01;
figure;
plot(Vn(:,p),Vn(:,p-1),'*');
title('Representation des variables avec deux l''axes de l''ACP');
xlabel('axe 1 : 11.74 % d''information');
ylabel('axe 2 : 7.69 % d''information');
hold on;
plot(cos(t),sin(t));
grid on;
for i=1:p % affichage du nom des variables sur le graphique
    text(Vn(i,p),Vn(i,p-1),Var_Name(i,:),'FontSize',9);
end;
%% Regression lineaire simple
%% G1 explicative G3 a expliquer
X=[Mat_G1 ones(n,1)];
[n,p]=size(X);
alpha_G1=X\Mat_G3;% matrice contenant les parametres de la regression
y_pred_G1=X*alpha_G1;% valeurs predites par la regression
residus_G1=Mat_G3-y_pred_G1;% erreurs entre les valeurs predites et les valeurs reelles
H = X*inv(X'*X)*X';% projecteur
h=diag(H);
variance_estimee = residus_G1'*residus_G1/(n-p-1);
residus_standardises = residus_G1./sqrt(variance_estimee*(1-h));
contributions_G1 = residus_standardises.*residus_standardises.*(h./(p.*(1-h)));
R2_G1=sum((y_pred_G1-mean(Mat_G3)).^2)/sum((Mat_G3-mean(Mat_G3)).^2);% coefficient de
    determination
figure;
plot(Mat_G1,y_pred_G1,'r');
title('Regression lineaire');
xlabel('Premieres notes');
ylabel('Notes finales');
hold on;
plot(Mat_G1,Mat_G3,'o');
figure;
```

```
plot (Mat_G3,residus_G1,'+r');
title('Residus');
xlabel('Notes finales');
ylabel('Modalites');
figure;
plot(Mat_G3,contributions_G1,'+r');
title('Contributions');
xlabel('Notes finales');
ylabel('Modalites');
% elimination des point aberrants c'est a dire des notes egales a 0
indices_aberrants=find(Mat_G3==0);
Mat_G1stock=Mat_G1;
Mat_G1(indices_aberrants)=[];
Mat_G3stock=Mat_G3;
Mat_G3(indices_aberrants)=[];
[n,p]=size(Mat_G1);
X=[Mat_G1 ones(n,1)];
[n,p]=size(X);
alpha_G1bis=X\Mat_G3;
y_pred_G1bis=X*alpha_G1bis;
residus_G1bis=Mat_G3-y_pred_G1bis;
H = X*inv(X'*X)*X';
h=diag(H);
variance_estimee = residus_G1bis'*residus_G1bis/(n-p-1);
residus_standardises = residus_G1bis./sqrt(variance_estimee*(1-h));
contributions_G1bis = residus_standardises.*residus_standardises.*(h./(p.*(1-h)));
R2_G1bis=sum((y_pred_G1bis-mean(Mat_G3)).^2)/sum((Mat_G3-mean(Mat_G3)).^2);
figure;
plot(Mat_G1,y_pred_G1bis,'r');
title('Regression lineaire');
xlabel('Premieres notes');
ylabel('Notes finales');
hold on;
plot(Mat_G1,Mat_G3,'o');
figure;
plot (Mat_G3,residus_G1bis,'o');
title ('Residus');
xlabel('Notes finales');
ylabel('Modalites');
[modalites N]=unique(sort(residus_G1bis));
effectifs=[diff(N);1];
figure;
bar(modalites,effectifs);
title('Distribution des residus');
xlabel('Modalites');
ylabel('Effectifs');
figure;
plot(Mat_G3,contributions_G1bis,'o');
title('Contributions');
xlabel('Notes finales');
ylabel('Modalites');

%% G2 explicative G3 a expliquer
Mat_G3=Mat_G3stock;
[n,p]=size(Mat_G2);
X=[Mat_G2 ones(n,1)];
[n,p]=size(X);
alpha_G2=X\Mat_G3;
y_pred_G2=X*alpha_G2;
residus_G2=Mat_G3-y_pred_G2;
H = X*inv(X'*X)*X';
```

```
168 h=diag(H);
169 variance_estimee = residus_G2'*residus_G2/(n-p-1);
170 residus_standardises = residus_G2./sqrt(variance_estimee*(1-h));
171 contributions_G2 = residus_standardises.*residus_standardises.*(h./(p.*(1-h)));
172 R2_G2=sum((y_pred_G2-mean(Mat_G3)).^2)/sum((Mat_G3-mean(Mat_G3)).^2);
173 figure;
174 plot(Mat_G2,y_pred_G2,'r');
175 title('Regression lineaire');
176 xlabel('Deuxiemes notes');
177 ylabel('Notes finales');
178 hold on;
179 plot(Mat_G2,Mat_G3,'o');
180 figure;
181 plot (Mat_G3,residus_G2,'+r');
182 title('Residus');
183 xlabel('Notes finales');
184 ylabel('Modalites');
185 figure;
186 plot(Mat_G3,contributions_G2,'o');
187 title('Contributions');
188 xlabel('Notes finales');
189 ylabel('Modalites');
190 % elimination des point aberrants c'est a dire des notes egales a 0
191 Mat_G2stock=Mat_G2;
192 Mat_G2(indices_aberrants)=[];
193 Mat_G3(indices_aberrants)=[];
194 [n,p]=size(Mat_G2);
195 X=[Mat_G2 ones(n,1)];
196 [n,p]=size(X);
197 alpha_G2bis=X\Mat_G3;
198 y_pred_G2bis=X*alpha_G2bis;
199 residus_G2bis=Mat_G3-y_pred_G2bis;
200 H = X*inv(X'*X)*X';
201 h=diag(H);
202 variance_estimee = residus_G2bis'*residus_G2bis/(n-p-1);
203 residus_standardises = residus_G2bis./sqrt(variance_estimee*(1-h));
204 contributions_G2bis = residus_standardises.*residus_standardises.*(h./(p.*(1-h)));
205 R2_G2bis=sum((y_pred_G2bis-mean(Mat_G3)).^2)/sum((Mat_G3-mean(Mat_G3)).^2);
206 figure;
207 plot(Mat_G2,y_pred_G2bis,'r');
208 title('Regression lineaire');
209 xlabel('Deuxiemes notes');
210 ylabel('Notes finales');
211 hold on;
212 plot(Mat_G2,Mat_G3,'o');
213 figure;
214 plot (Mat_G3,residus_G2bis,'+r');
215 title('Residus')
216 xlabel('Notes finales');
217 ylabel('Modalites');
218 [modalites N]=unique(sort(residus_G2bis));
219 effectifs=[diff(N);2];
220 figure;
221 bar(modalites,effectifs);
222 title('Distribution des residus');
223 xlabel('Modalites');
224 ylabel('Effectifs');
225 figure;
226 plot(Mat_G3,contributions_G2bis,'o');
227 title('Contributions');
228 xlabel('Notes finales');
```

```
ylabel('Modalites');
%% Regression lineaire multiple
%% G1 et G2 explicatives G3 a expliquer
Mat_G1=Mat_G1stock;
Mat_G2=Mat_G2stock;
Mat_G3=Mat_G3stock;
[n p]=size(Mat_G3);
X=[Mat_G1 Mat_G2 ones(n,1)];
[n p]=size(X);
alpha=X\Mat_G3;
y_pred=X*alpha;
residus=Mat_G3-y_pred;
H = X*inv(X'*X)*X';
h=diag(H);
variance_estimee = residus'*residus/(n-p-1);
residus_standardises = residus./sqrt(variance_estimee*(1-h));
contributions= residus_standardises.*residus_standardises.*(h./(p.*(1-h)));
R2=sum((y_pred-mean(Mat_G3)).^2)/sum((Mat_G3-mean(Mat_G3)).^2);
figure;
plot (Mat_G3,residus,'+r');
title('Residus');
xlabel('Notes finales');
ylabel('Modalites');
figure;
plot(Mat_G3,contributions,'o');
title('Contributions');
xlabel('Notes finales');
ylabel('Modalites');
% elimination des point aberrants c'est a dire des notes egales a 0
Mat_G1(indices_aberrants)=[];
Mat_G2(indices_aberrants)=[];
Mat_G3(indices_aberrants)=[];
[n p]=size(Mat_G3);
X=[Mat_G1 Mat_G2 ones(n,1)];
[n p]=size(X);
alpha_bis=X\Mat_G3;
y_pred_bis=X*alpha_bis;
residus_bis=Mat_G3-y_pred_bis;
H = X*inv(X'*X)*X';
h=diag(H);
variance_estimee = residus_bis'*residus_bis/(n-p-1);
residus_standardises = residus_bis./sqrt(variance_estimee*(1-h));
contributions_bis= residus_standardises.*residus_standardises.*(h./(p.*(1-h)));
R2_bis=sum((y_pred_bis-mean(Mat_G3)).^2)/sum((Mat_G3-mean(Mat_G3)).^2);
figure;
plot (Mat_G3,residus_bis,'+r');
title('Residus');
xlabel('Notes finales');
ylabel('Modalites');
[modalites N]=unique(sort(residus_bis));
effectifs=[diff(N);1];
figure;
bar(modalites,effectifs);
title('Distribution des residus');
xlabel('Modalites');
ylabel('Effectifs');
figure;
plot(Mat_G3,contributions_bis,'o');
title('Contributions');
xlabel('Notes finales');
ylabel('Modalites');
```

```
290 %% Comparaison Failures note finale
291 Mat_G3=Mat_G3stock;
292 indice_0redoublement=find(Mat_Failures==0);
293 effectif_0redoublement=Mat_G3(indice_0redoublement);
294 moyenne_0redoublement=mean(effectif_0redoublement);
295 figure;
296 boxPlot(effectif_0redoublement);
297 title('Notes finales des eleves ayant eu 0 redoublement');
298 ylabel('Notes');
299 grid on;
300 indice_1redoublement=find(Mat_Failures==1);
301 effectif_1redoublement=Mat_G3(indice_1redoublement);
302 moyenne_1redoublement=mean(effectif_1redoublement);
303 figure;
304 boxPlot(effectif_1redoublement);
305 title('Notes finales des eleves ayant eu 1 redoublement');
306 ylabel('Notes');
307 grid on;
308 indice_2redoublements=find(Mat_Failures==2);
309 effectif_2redoublements=Mat_G3(indice_2redoublements);
310 moyenne_2redoublements=mean(effectif_2redoublements);
311 figure;
312 boxPlot(effectif_2redoublements);
313 title('Notes finales des eleves ayant eu 2 redoublements');
314 ylabel('Notes');
315 grid on;
316 indice_3redoublements=find(Mat_Failures==3);
317 effectif_3redoublements=Mat_G3(indice_3redoublements);
318 moyenne_3redoublements=mean(effectif_3redoublements);
319 figure;
320 boxPlot(effectif_3redoublements);
321 title('Notes finales des eleves ayant eu 3 redoublements');
322 ylabel('Notes');
323 grid on;
324 histogramme=[moyenne_0redoublement moyenne_1redoublement moyenne_2redoublements
        moyenne_3redoublements];
325 figure;
326 bar(histogramme);
327 title('Evolution des moyennes en fonction du nombre de redoublement des eleves');
328 ylabel('Moyenne des notes finales obtenues');
329 set(gca,'XTickLabel',{'0 redoublement','1 redoublement','2 redoublements','3
        redoublements'});
330 %% Comparaison Medu Note finale
331 indice_0degre_education_mere=find(Mat_Medu==0);
332 effectif_0degre_education_mere=Mat_G3(indice_0degre_education_mere);
333 indice_1degre_education_mere=find(Mat_Medu==1);
334 effectif_1degre_education_mere=Mat_G3(indice_1degre_education_mere);
335 moyenne_1degre_education_mere=mean(effectif_1degre_education_mere);
336 figure;
337 boxPlot(effectif_1degre_education_mere);
338 title('Notes finales des eleves ayant une mere avec un niveau 1 d''education');
339 ylabel('Notes');
340 grid on;
341 indice_2degre_education_mere=find(Mat_Medu==2);
342 effectif_2degre_education_mere=Mat_G3(indice_2degre_education_mere);
343 moyenne_2degre_education_mere=mean(effectif_2degre_education_mere);
344 figure;
345 boxPlot(effectif_2degre_education_mere);
346 title('Notes finales des eleves ayant une mere avec un niveau 2 d''education');
347 ylabel('Notes');
348 grid on;
```

```
indice_3degre_education_mere=find(Mat_Medu==3);
effectif_3degre_education_mere=Mat_G3(indice_3degre_education_mere);
moyenne_3degre_education_mere=mean(effectif_3degre_education_mere);
figure;
boxPlot(effectif_3degre_education_mere);
title('Notes finales des eleves ayant une mere avec un niveau 3 d''education');
ylabel('Notes');
grid on;
indice_4degre_education_mere=find(Mat_Medu==4);
effectif_4degre_education_mere=Mat_G3(indice_4degre_education_mere);
moyenne_4degre_education_mere=mean(effectif_4degre_education_mere);
figure;
boxPlot(effectif_4degre_education_mere);
title('Notes finales des eleves ayant une mere avec un niveau 4 d''education');
ylabel('Notes');
grid on;
histogramme=[moyenne_1degre_education_mere moyenne_2degre_education_mere
    moyenne_3degre_education_mere moyenne_4degre_education_mere];
figure;
bar(histogramme);
title('Evolution des moyennes en fonction du niveau d''education des meres des eleves');
ylabel('Moyenne des notes finales obtenues');
set(gca,'XTickLabel',{'Niveau 1','Niveau 2','Niveau 3','Niveau 4'});
%% Comparaison Fedu Note finale
indice_0degre_education_pere=find(Mat_Fedu==0);
effectif_0degre_education_pere=Mat_G3(indice_0degre_education_pere);
indice_1degre_education_pere=find(Mat_Fedu==1);
effectif_1degre_education_pere=Mat_G3(indice_1degre_education_pere);
moyenne_1degre_education_pere=mean(effectif_1degre_education_pere);
figure;
boxPlot(effectif_1degre_education_pere);
title('Notes finales des eleves ayant un pere avec un niveau 1 d''education');
ylabel('Notes');
grid on;
indice_2degre_education_pere=find(Mat_Fedu==2);
effectif_2degre_education_pere=Mat_G3(indice_2degre_education_pere);
moyenne_2degre_education_pere=mean(effectif_2degre_education_pere);
figure;
boxPlot(effectif_2degre_education_pere);
title('Notes finales des eleves ayant un pere avec un niveau 2 d''education');
ylabel('Notes');
grid on;
indice_3degre_education_pere=find(Mat_Fedu==3);
effectif_3degre_education_pere=Mat_G3(indice_3degre_education_pere);
moyenne_3degre_education_pere=mean(effectif_3degre_education_pere);
figure;
boxPlot(effectif_3degre_education_pere);
title('Notes finales des eleves ayant un pere avec un niveau 3 d''education');
ylabel('Notes');
grid on;
indice_4degre_education_pere=find(Mat_Fedu==4);
effectif_4degre_education_pere=Mat_G3(indice_4degre_education_pere);
moyenne_4degre_education_pere=mean(effectif_4degre_education_pere);
figure;
boxPlot(effectif_4degre_education_pere);
title('Notes finales des eleves ayant un pere avec un niveau 4 d''education');
ylabel('Notes');
grid on;
histogramme=[moyenne_1degre_education_pere moyenne_2degre_education_pere
    moyenne_3degre_education_pere moyenne_4degre_education_pere];
figure;
```

```
bar(histogramme);
title('Evolution des moyennes en fonction du niveau d''education des peres des eleves');
ylabel('Moyenne des notes finales obtenues');
set(gca,'XTickLabel',{'Niveau 1','Niveau 2','Niveau 3','Niveau 4'});
%% Comparaison Age note finale
indice_15ans=find(Mat_Age==15);
effectif_15ans=Mat_G3(indice_15ans);
moyenne_15ans=mean(effectif_15ans);
figure;
boxPlot(effectif_15ans);
title('Notes finales des eleves ayant 15 ans');
ylabel('Notes');
grid on;
indice_16ans=find(Mat_Age==16);
effectif_16ans=Mat_G3(indice_16ans);
moyenne_16ans=mean(effectif_16ans);
figure;
boxPlot(effectif_16ans);
title('Notes finales des eleves ayant 16 ans');
ylabel('Notes');
grid on;
indice_17ans=find(Mat_Age==17);
effectif_17ans=Mat_G3(indice_17ans);
moyenne_17ans=mean(effectif_17ans);
figure;
boxPlot(effectif_17ans);
title('Notes finales des eleves ayant 17 ans');
ylabel('Notes');
grid on;
indice_18ans=find(Mat_Age==18);
effectif_18ans=Mat_G3(indice_18ans);
moyenne_18ans=mean(effectif_18ans);
figure;
boxPlot(effectif_18ans);
title('Notes finales des eleves ayant 18 ans');
ylabel('Notes');
grid on;
indice_19ans=find(Mat_Age==19);
effectif_19ans=Mat_G3(indice_19ans);
moyenne_19ans=mean(effectif_19ans);
figure;
boxPlot(effectif_19ans);
title('Notes finales des eleves ayant 19 ans');
ylabel('Notes');
grid on;
indice_20ans=find(Mat_Age==20);
effectif_20ans=Mat_G3(indice_20ans);
indice_21ans=find(Mat_Age==21);
effectif_21ans=Mat_G3(indice_21ans);
indice_22ans=find(Mat_Age==22);
effectif_22ans=Mat_G3(indice_22ans);
histogramme=[moyenne_15ans moyenne_16ans moyenne_17ans moyenne_18ans moyenne_19ans];
figure;
bar(histogramme,'r');
title('Evolution des moyennes en fonction de l''age des eleves');
ylabel('Moyenne des notes finales obtenues');
set(gca,'XTickLabel',{'15 ans','16 ans','17 ans','18 ans','19 ans'});
```

Annexe C – `Test_Chi2.m`

```
%run('DonneesProjetM8.m')

%% Test du Chi2

    %On decide ici de tester l'independance des deux variables 'internet'
    %et 'studytime'

    %% Etape 1 : construire le tableau de contingence
        %H0 : 'Internet' et 'Studytime' sont independantes
        %H1 : 'Internet' et 'Studytime' sont liees

    % Tout d'abord, on construit le tableau de contingence O des
    % observations (2 variables qualitatives de resp. I et J modalites)

    temp=[TabModMat_Internet , Mat_Studytime];

    O = zeros(2,4); % 2 lignes pour 'internet' et 4 colonnes pour 'studytime'

    ind=[]; %Pour [0,1]
    for i=1:length(temp)
        if temp(i,:)==[0,1]
            ind=[ind i];
        end
    end
   O(1,1)=length(ind);

    ind=[]; %Pour [0,2]
    for i=1:length(temp)
        if temp(i,:)==[0,2]
            ind=[ind i];
        end
    end
   O(1,2)=length(ind);

    ind=[]; %Pour [0,3]
    for i=1:length(temp)
        if temp(i,:)==[0,3]
            ind=[ind i];
        end
    end
   O(1,3)=length(ind);

    ind=[]; %Pour [0,4]
    for i=1:length(temp)
        if temp(i,:)==[0,4]
            ind=[ind i];
        end
    end
   O(1,4)=length(ind);

    ind=[]; %Pour [1,1]
    for i=1:length(temp)
        if temp(i,:)==[1,1]
            ind=[ind i];
        end
    end
   O(2,1)=length(ind);

```

```
59      ind=[]; %Pour [1,2]
60      for i=1:length(temp)
61          if temp(i,:)==[1,2]
62              ind=[ind i];
63          end
64      end
65     O(2,2)=length(ind);
66
67      ind=[]; %Pour [1,3]
68      for i=1:length(temp)
69          if temp(i,:)==[1,3]
70              ind=[ind i];
71          end
72      end
73     O(2,3)=length(ind);
74
75      ind=[]; %Pour [1,4]
76      for i=1:length(temp)
77          if temp(i,:)==[1,4]
78              ind=[ind i];
79          end
80      end
81     O(2,4)=length(ind);
82     clear i;
83     clear ind;
84
85     %% Etape 2 : on calcule les marginales
86      [I,J]=size(O); %[2,4]
87      nI=sum(O'); %profil ligne = [66 329]
88      nJ=sum(O); %profil colonne = [105 198 65 27]
89      n=sum(sum(O)); % = 395
90
91     %% Etape 3 : on calcule lees Tij pour chaque case du tab des eff theoriques
92      T=(nI'*nJ)/n; % Effectifs theoriques <==> Tij=P(i.)*P(.j)*n
93                                            %= Ni./n * N.j/n * n
94      % T/n donne les pourcentages
95
96     %% Etape 4 : on calcule la distance du Chi2 :
97      D= sum(sum((O-T).^2./T)); %Distance du Chi2 : 3.3831
98
99     %% Etape 5 : on calcule le nombre de degres de liberte
100     ddl=(I-1)*(J-1); %3
101
102    %% Etape 6 : on regarde dans les tables de la loi de Chi2
103     %pval=1-chi2cdf(3.3831,3) %Cf site 'octave-online.net'
104     %pval =  0.33624
105
106    %% Etape 7 : conclure
107    % On constate que pval<0.05
108    % On garde donc H0 : il n'y a pas lieu de remettre en cause
109    % l'independance des variables 'Internet' et 'Studytime'
110
111
112
113
114
115
116
117
118 %% Test du Chi2 (Romantic / Walc)
119
```

```
120      %On decide ici de tester l'independance des deux variables 'romantic'
121      %et 'walc'
122
123      %% Etape 1 : construire le tableau de contingence
124          %H0 : 'Romantic' et 'Walc' sont independantes
125          %H1 : 'Romantic' et 'Walc' sont liees
126
127      % Tout d'abord, on construit le tableau de contingence O des
128      % observations (2 variables qualitatives de resp. I et J modalites)
129
130      temp=[TabModMat_Romantic , Mat_Walc];
131
132      O = zeros(2,5); % 2 lignes pour 'romantic' et 4 colonnes pour 'Walc'
133
134      ind=[]; %Pour [0,1]
135      for i=1:length(temp)
136          if temp(i,:)==[0,1]
137              ind=[ind i];
138          end
139      end
140     O(1,1)=length(ind);
141
142      ind=[]; %Pour [0,2]
143      for i=1:length(temp)
144          if temp(i,:)==[0,2]
145              ind=[ind i];
146          end
147      end
148     O(1,2)=length(ind);
149
150      ind=[]; %Pour [0,3]
151      for i=1:length(temp)
152          if temp(i,:)==[0,3]
153              ind=[ind i];
154          end
155      end
156     O(1,3)=length(ind);
157
158      ind=[]; %Pour [0,4]
159      for i=1:length(temp)
160          if temp(i,:)==[0,4]
161              ind=[ind i];
162          end
163      end
164     O(1,4)=length(ind);
165
166         ind=[]; %Pour [0,5]
167      for i=1:length(temp)
168          if temp(i,:)==[0,5]
169              ind=[ind i];
170          end
171      end
172     O(1,5)=length(ind);
173
174      ind=[]; %Pour [1,1]
175      for i=1:length(temp)
176          if temp(i,:)==[1,1]
177              ind=[ind i];
178          end
179      end
180     O(2,1)=length(ind);
```

```
181
182      ind=[]; %Pour [1,2]
183      for i=1:length(temp)
184          if temp(i,:)==[1,2]
185              ind=[ind i];
186          end
187      end
188     O(2,2)=length(ind);
189
190      ind=[]; %Pour [1,3]
191      for i=1:length(temp)
192          if temp(i,:)==[1,3]
193              ind=[ind i];
194          end
195      end
196     O(2,3)=length(ind);
197
198      ind=[]; %Pour [1,4]
199      for i=1:length(temp)
200          if temp(i,:)==[1,4]
201              ind=[ind i];
202          end
203      end
204     O(2,4)=length(ind);
205
206      ind=[]; %Pour [1,5]
207      for i=1:length(temp)
208          if temp(i,:)==[1,5]
209              ind=[ind i];
210          end
211      end
212     O(2,5)=length(ind);
213     clear i;
214     clear ind;
215
216     %% Etape 2 : on calcule les marginales
217      [I,J]=size(O); %[2,5]
218      nI=sum(O'); %profilligne = [263 132]
219      nJ=sum(O); %profil colonne = [151 85 80 51 28]
220      n=sum(sum(O)); % = 395
221
222     %% Etape 3 : on calcule lees Tij pour chaque case du tab des eff theoriques
223      T=(nI'*nJ)/n; % Effectifs theoriques <==> Tij=P(i.)*P(.j)*n
224                                            %= Ni./n * N.j/n * n
225      % T/n donne les pourcentages
226
227     %% Etape 4 : on calcule la distance du Chi2 :
228      D= sum(sum((O-T).^2./T)); %Distance du Chi2 : 1.9912
229
230     %% Etape 5 : on calcule le nombre de degres de liberte
231      ddl=(I-1)*(J-1); %4
232
233     %% Etape 6 : on regarde dans les tables de la loi de Chi2
234      %pval=1-chi2cdf(1,9912,4) %Cf site 'octave-online.net'
235      %pval =  0.73738
236
237     %% Etape 7 : conclure
238     % On constate que pval>0.05
239     % On ne peut donc pas conclure quant a la dependance des deux variables
240
241
```

```
242
243
244
245
246     %% Test du Chi2 (Medu / Fedu)
247
248      %On decide ici de tester l'independance des deux variables 'Medu'
249      %et 'Fedu'
250
251      %% Etape 1 : construire le tableau de contingence
252          %H0 : 'Medu' et 'Fedu' sont independantes
253          %H1 : 'Medu' et 'Fedu' sont liees
254
255      % Tout d'abord, on construit le tableau de contingence O des
256      % observations (2 variables qualitatives de resp. I et J modalites)
257
258      temp=[Mat_Medu , Mat_Fedu];
259
260      O = zeros(5,5); % 5 lignes pour 'Medu' et 5 colonnes pour 'Fedu'
261
262      for j=0:4
263          ind=[]; %Pour [0,i]
264          for i=1:length(temp)
265              if temp(i,:)==[0,j]
266                  ind=[ind i];
267              end
268          end
269          O(1,j+1)=length(ind);
270      end
271
272
273      for j=0:4
274          ind=[]; %Pour [1,i]
275          for i=1:length(temp)
276              if temp(i,:)==[1,j]
277                  ind=[ind i];
278              end
279          end
280          O(2,j+1)=length(ind);
281      end
282
283      for j=0:4
284          ind=[]; %Pour [2,i]
285          for i=1:length(temp)
286              if temp(i,:)==[2,j]
287                  ind=[ind i];
288              end
289          end
290          O(3,j+1)=length(ind);
291      end
292
293      for j=0:4
294          ind=[]; %Pour [3,i]
295          for i=1:length(temp)
296              if temp(i,:)==[3,j]
297                  ind=[ind i];
298              end
299          end
300          O(4,j+1)=length(ind);
301      end
302
```

```
303      for j=0:4
304          ind=[]; %Pour [4,i]
305          for i=1:length(temp)
306              if temp(i,:)==[4,j]
307                  ind=[ind i];
308              end
309          end
310          O(5,j+1)=length(ind);
311      end
312
313
314     clear i;
315     clear ind;
316
317     %% Etape 2 : on calcule les marginales
318      [I,J]=size(O); %[5,5]
319      nI=sum(O'); %profil ligne = [3 59 103 99 131]
320      nJ=sum(O); %profil colonne = [2 82 115 100 96]
321      n=sum(sum(O)); % = 395
322
323     %% Etape 3 : on calcule lees Tij pour chaque case du tab des eff theoriques
324      T=(nI'*nJ)/n; % Effectifs theoriques <==> Tij=P(i.)*P(.j)*n
325                                            %= Ni./n * N.j/n * n
326      % T/n donne les pourcentages
327
328     %% Etape 4 : on calcule la distance du Chi2 :
329      D= sum(sum((O-T).^2./T)); %Distance du Chi2 : 199.9773
330
331     %% Etape 5 : on calcule le nombre de degres de liberte
332      ddl=(I-1)*(J-1); %16
333
334     %% Etape 6 : on regarde dans les tables de la loi de Chi2
335      %pval=1-chi2cdf(199.9774,16) %Cf site 'octave-online.net'
336      %pval =  0
337
338     %% Etape 7 : conclure
339     % On constate que pval<0.05
340     % On garde donc H0 : il n'y a pas lieu de remettre en cause
341     % l'independance des variables 'Medu' et 'Fedu'
342
343
344
345
346
347
348        %% Test du Chi2 (Mjob / Fjob)
349
350      %On decide ici de tester l'independance des deux variables 'Mjob'
351      %et 'Fjob'
352
353      %% Etape 1 : construire le tableau de contingence
354          %H0 : 'Mjob' et 'Fjob' sont independantes
355          %H1 : 'Mjob' et 'Fjob' sont liees
356
357      % Tout d'abord, on construit le tableau de contingence O des
358      % observations (2 variables qualitatives de resp. I et J modalites)
359
360      temp=[TabModMat_Mjob , TabModMat_Fjob];
361
362      O = zeros(5,5); % 5 lignes pour 'Mjob' et 5 colonnes pour 'Fjob'
363
```

```
364      for j=0:4
365          ind=[]; %Pour [0,i]
366          for i=1:length(temp)
367              if temp(i,:)==[0,j]
368                  ind=[ind i];
369              end
370          end
371          O(1,j+1)=length(ind);
372      end
373
374
375      for j=0:4
376          ind=[]; %Pour [1,i]
377          for i=1:length(temp)
378              if temp(i,:)==[1,j]
379                  ind=[ind i];
380              end
381          end
382          O(2,j+1)=length(ind);
383      end
384
385      for j=0:4
386          ind=[]; %Pour [2,i]
387          for i=1:length(temp)
388              if temp(i,:)==[2,j]
389                  ind=[ind i];
390              end
391          end
392          O(3,j+1)=length(ind);
393      end
394
395      for j=0:4
396          ind=[]; %Pour [3,i]
397          for i=1:length(temp)
398              if temp(i,:)==[3,j]
399                  ind=[ind i];
400              end
401          end
402          O(4,j+1)=length(ind);
403      end
404
405      for j=0:4
406          ind=[]; %Pour [4,i]
407          for i=1:length(temp)
408              if temp(i,:)==[4,j]
409                  ind=[ind i];
410              end
411          end
412          O(5,j+1)=length(ind);
413      end
414
415
416     clear i;
417     clear ind;
418
419     %% Etape 2 : on calcule les marginales
420      [I,J]=size(O); %[5,5]
421      nI=sum(O'); %profil ligne = [59 34 141 103 58]
422      nJ=sum(O); %profil colonne = [20 18 217 111 29]
423      n=sum(sum(O)); % = 395
424
```

```
%% Etape 3 : on calcule lees Tij pour chaque case du tab des eff theoriques
 T=(nI'*nJ)/n; % Effectifs theoriques <==> Tij=P(i.)*P(.j)*n
                                       %= Ni./n * N.j/n * n
 % T/n donne les pourcentages

%% Etape 4 : on calcule la distance du Chi2 :
 D= sum(sum((O-T).^2./T)); %Distance du Chi2 : 73.3809

%% Etape 5 : on calcule le nombre de degres de liberte
 ddl=(I-1)*(J-1); %16

%% Etape 6 : on regarde dans les tables de la loi de Chi2
 %pval=1-chi2cdf(73.3809,16) %Cf site 'octave-online.net'
 %pval =  2.5336e-09

%% Etape 7 : conclure
% On constate que pval<0.05
% On garde donc H0 : il n'y a pas lieu de remettre en cause
% l'independance des variables 'Mjob' et 'Fjob'









      %% Test du Chi2 (Dalc / Walc)

 %On decide ici de tester l'independance des deux variables 'Dalc'
 %et 'Walc'

 %% Etape 1 : construire le tableau de contingence
     %H0 : 'Dalc' et 'Walc' sont independantes
     %H1 : 'Dalc' et 'Walc' sont liees

 % Tout d'abord, on construit le tableau de contingence O des
 % observations (2 variables qualitatives de resp. I et J modalites)

 temp=[Mat_Dalc , Mat_Walc];

 O = zeros(5,5); % 5 lignes pour 'Dalc' et 5 colonnes pour 'Walc'

 for j=1:5
     ind=[]; %Pour [1,i]
     for i=1:length(temp)
         if temp(i,:)==[1,j]
             ind=[ind i];
         end
     end
     O(1,j)=length(ind);
 end


 for j=1:5
     ind=[]; %Pour [2,i]
     for i=1:length(temp)
         if temp(i,:)==[2,j]
             ind=[ind i];
         end
     end
     O(2,j)=length(ind);
 end

```

```
    for j=1:5
        ind=[]; %Pour [3,i]
        for i=1:length(temp)
            if temp(i,:)==[3,j]
                ind=[ind i];
            end
        end
        O(3,j)=length(ind);
    end

    for j=1:5
        ind=[]; %Pour [4,i]
        for i=1:length(temp)
            if temp(i,:)==[4,j]
                ind=[ind i];
            end
        end
        O(4,j)=length(ind);
    end

    for j=1:5
        ind=[]; %Pour [5,i]
        for i=1:length(temp)
            if temp(i,:)==[5,j]
                ind=[ind i];
            end
        end
        O(5,j)=length(ind);
    end


clear i;
clear ind;

%% Etape 2 : on calcule les marginales
 [I,J]=size(O); %[5,5]
 nI=sum(O'); %profil ligne = [276 75 26 9 9]
 nJ=sum(O); %profil colonne = [151 85 80 51 28]
 n=sum(sum(O)); % = 395

%% Etape 3 : on calcule lees Tij pour chaque case du tab des eff theoriques
 T=(nI'*nJ)/n; % Effectifs theoriques <==> Tij=P(i.)*P(.j)*n
                                        %= Ni./n * N.j/n * n
 % T/n donne les pourcentages

%% Etape 4 : on calcule la distance du Chi2 :
 D= sum(sum((O-T).^2./T)); %Distance du Chi2 : 287.0019

%% Etape 5 : on calcule le nombre de degres de liberte
 ddl=(I-1)*(J-1); %16

%% Etape 6 : on regarde dans les tables de la loi de Chi2
 %pval=1-chi2cdf(287.0019,16) %Cf site 'octave-online.net'
 %pval =  0

%% Etape 7 : conclure
% On constate que pval<0.05
% On garde donc H0 : il n'y a pas lieu de remettre en cause
% l'independance des variables 'Dalc' et 'Walc'
```

Annexe D – `Test_Student.m`

```
%run('DonneesProjetM8.m')

%% Test de Student
% Il semble coherent de faire un test de Student sur des variables
% quantitatives qui representent reellement une quantite.
% Ainsi, 'age' et 'absences' seront soumises au test de Student


        %% Etape 1 : Formuler les hypotheses

            %H0 : l'age de l'eleve n'est pas en lien avec son taux d'absenteisme
            %H1 : l'age de l'eleve est lie a son taux d'absenteisme



        %% Etape 2 : Poser un modele
            %Soit  Xb_age (Xb : x-barre) la v.a. representant la moyenne des ages
            %Xb_age~N(μ_age,sigma2_age/n)

            Xb_age=mean(Mat_Age); % 16.6962 : age moyen

            %Soit  Xb_absences (Xb : x-barre) la v.a. representant la moyenne des
            %absences
            %Xb_absences~N(μ_absences,sigma2_absences/n)

            Xb_absences=mean(Mat_Absences); % 5.7089 : absenteisme moyen

            %Calcul du Sigma2 :
            sigma2=(1/(length(Mat_Absences) + length(Mat_Age)-2));
            sigma2=sigma2*(sum((Mat_Absences-Xb_absences).^2)+sum((Mat_Age-Xb_age).^2));
                %% 32.8389


            %Reformuler les hypotheses :
                %H0 : μ_age = μ_absences
                %H1 : μ_age <> μ_absences


        %% Etape 3 : Exhiber la statistique du test
            t=(Xb_age-Xb_absences) / (sqrt(sigma2*(1/length(Mat_Age) + 1/length(
                Mat_Absences)))); %26.9452
            %Nombre de degres de liberte :
            ddl = length(Mat_Age)+length(Mat_Absences)-2; %788

        %% Etape 4 :Calculer la P-Valeur
            % On calcule la P-Valeur :
            % P_val=1-cdf('t',26.9452, 788) ; %cf site "octave-online.net" pour cdf

            % P_val = 0
        %% Etape 5 : Conclure
            %la p_val etant inferieure a alpha=0.05,
                %on garde H0
                %L'age des eleves n'est pas en lien avec leur taux d'absenteisme

                %Ce resultat etait previsible et ce test n'a que peu d'interet.
```

```
57
58
59
60
61          %% Etape 1 : Formuler les hypotheses
62
63              %H0 : l'age de l'eleve n'est pas en lien avec son taux d'absenteisme
64              %H1 : l'age de l'eleve est lie a son taux d'absenteisme
65
66
67
68          %% Etape 2 : Poser un modele
69              %Soit  Xb_age (Xb : x-barre) la v.a. representant la moyenne des ages
70              %Xb_age~N(μ_age,sigma2_age/n)
71
72              Xb_ageP=mean(Por_Age); % 16.7442 : age moyen
73
74              %Soit  Xb_absences (Xb : x-barre) la v.a. representant la moyenne des
75              %absences
76              %Xb_absences~N(μ_absences,sigma2_absences/n)
77
78              Xb_absencesP=mean(Por_Absences); % 3.6595 : absenteisme moyen
79
80              %Calcul du Sigma2 :
81              sigma2P=(1/(length(Por_Absences) + length(Por_Age)-2));
82              sigma2P=sigma2P*(sum((Por_Absences-Xb_absencesP).^2)+sum((Por_Age-Xb_ageP).
                    ^2)); %% 11.5103
83
84
85              %Reformuler les hypotheses :
86                  %H0 : μ_age = μ_absences
87                  %H1 : μ_age <> μ_absences
88
89
90          %% Etape 3 : Exhiber la statistique du test
91              tP=(Xb_ageP-Xb_absencesP) / (sqrt(sigma2P*(1/length(Por_Age) + 1/length(
                    Por_Absences)))); %69.4753
92              %Nombre de degres de liberte :
93              ddlP = length(Por_Age)+length(Por_Absences)-2; %1296
94
95          %% Etape 4 :Calculer la P-Valeur
96              % On calcule la P-Valeur :
97              % P_val=1-cdf('t',69.4753, 1296) ; %cf site "octave-online.net" pour cdf
98
99              % P_val = 0
100         %% Etape 5 : Conclure
101             %la p_val etant inferieure a alpha=0.05,
102                 %on garde H0
103                 %L'age des eleves n'est pas en lien avec leur taux d'absenteisme
104
105                 %Ce resultat etait previsible et ce test n'a que peu d'interet.
```