ANNÉE UNIVERSITAIRE 2022 – 2023

MASTER 2 ÉCONOMIE APPLIQUÉE PARCOURS INGÉNIERIE DES DONNÉES ET ÉVALUATIONS ÉCONOMÉTRIQUES

DOSSIER COMMUN

ATELIER MÉTHODES D'ÉVALUATIONS ÉCONOMIQUE & ATELIER DISCRIMINATION ET MARCHÉ DU TRAVAIL

ENSEIGNANT : M. CHRISTOPHE DANIEL

LAMON Océane - PEDROT Emma - SEZESTRE Emilien

université angers

Table des matières

I. Introduction

Le marché du travail est le lieu théorique de rencontre entre l'offre de travail -représentée par les individus qui proposent leur savoir-faire et compétences contre une rémunération-, et la demande de travail qui émane des entreprises privées ou publiques, voire de personnes privées dans certains cas. Une fois rémunéré, le travail peut être associé à un emploi du moment qu'il est juridiquement encadré, conformément à l'ensemble de lois qui composent le droit du travail. L'offreur et le demandeur de travail sont donc liés par un contrat de travail, qui fixe les droits et les devoirs de chaque partie prenante dans une temporalité prédéfinie ou indéfinie.

La rencontre entre offre et demande de travail doit être rendue possible par la communication d'une offre d'emploi sur un site spécialisé ou sur un réseau social pour mettre en relation les parties prenantes. Dans le cas le plus courant, le demandeur de travail rédige une offre d'emploi dans laquelle il spécifie les principales caractéristiques inhérentes au poste à pourvoir : type de contrat et durée, lieu, rémunération, principales missions et pré-requis en termes de formation ou d'expérience. Une fois l'annonce diffusée, les offreurs de travail intéressés par cette dernière vont déposer une candidature en présentant entre-autres leur formation, leurs compétences et leur expérience professionnelle. S'opère alors une évaluation des profils et une sélection au cours de laquelle le recruteur procède souvent à des entretiens pour échanger de vive voix avec les candidats, et ainsi arrêter son choix sur le plus convaincant d'entre eux.

Si l'on se place sur un premier axe d'analyse, il est commun de constater la présence de diverses formes de discriminations sur ce marché. Par discrimination, l'on entend les conséquences d'une "différenciation objectivement injustifiée" (Discrimination — Wikipédia (wikipedia.org)) généralement basée sur le genre, l'âge, la religion, le positionnement politique ou encore la présence de formes de handicap. Parmi les plus répandues, l'on peut citer la discrimination à l'embauche (assez difficile à mesurer car les motifs discriminatoires ne seront jamais explicitement évoqués lors du refus), la discrimination salariale (généralement mesurée entre hommes et femmes à expérience équivalente) ou encore le plafond de verre qui est une forme de discrimination associée à une inaccessibilité injustifiée des plus hautes sphères de la hiérarchie pour certaines catégories de personnes.

D'autre part, nous pouvons également mettre en évidence l'existence de politiques pour stimuler l'emploi. Parmi ces dernières, nous pouvons par exemple citer la mise en place d'un salaire plancher ou la modification de son seuil, des aides au retour à l'emploi, des impacts sur le montant de l'indemnité maladie en fonction de la durée de l'arrêt, etc. Il est cependant difficile d'attribuer un quelconque impact à la mise en place d'une mesure si on ne cherche pas à isoler les conséquences de cette dernière d'autres facteurs exogènes qui pourraient influer sur la situation.

Il est néanmoins possible de démontrer l'existence de discriminations et de quantifier leur effet moyen, ainsi que de vérifier l'effet significatif d'une politique mise en place sur le marché du travail, à l'aide de diverses méthodes d'évaluations économiques appropriées. Ces différents outils d'analyse ont constitué le cœur des deux ateliers Méthodes d'évaluations économiques appliquées au marché du travail & Discriminations et marché du travail. L'objectif de ce dossier étant de réutiliser les méthodes sollicitées en cours sur nos propres données, nous avons de ce fait pris la décision d'explorer un autre secteur : l'éducation. Ce dernier se prête tout aussi bien aux deux axes d'analyse évoqués précédemment, dans la mesure où il est envisageable de considérer l'existence de discriminations en milieu scolaire (notamment en phases d'admission dans le supérieur pour certaines filières) autant que la présence de mesures destinées à améliorer le niveau des élèves.

II. Présentation des données

Cette seconde partie a vocation à présenter les données qui constituent le fondement de notre étude, en passant par leur source originelle, les éventuelles transformations et modifications qui leur ont été apportées ainsi que la présentation des principales variables d'intérêt.

A. Source des données

Les données mobilisées dans le cadre de cette analyse proviennent à l'origine de l'étude *Using Data Mining to Predict Secondary School Student Performance* (2008) réalisée par Paulo Cortez et Alice Silva.

Pour prédire la note finale de lycéens (15-18 ans) et identifier les facteurs clés de la réussite scolaire, les auteurs se sont basés sur des données récoltées entre 2005 et 2006 dans deux écoles publiques portugaises situées dans la région d'Alentejo, par le biais de bulletins scolaires et de questionnaires administrés aux élèves. Après avoir fait valider leur questionnaire auprès du corps enseignant et l'avoir fait tester par un petit groupe d'étudiants, le questionnaire comportant 37 questions a été diffusé à l'échelle des deux établissements et, parmi les 788 répondants, seules les réponses de 677 d'entre eux ont été conservées, les autres ayant été rejetées car les données d'identification recueillies n'ont pas permis de les associer aux résultats scolaires extraits des bulletins de notes.

Enfin, les données ainsi recueillies ont été séparées en deux bases de données distinctes qui concernent spécifiquement l'enseignement de la langue portugaise ainsi que les mathématiques. Dans le cadre de cette étude, nous avons décidé de regrouper les données d'origine dans une unique base, et de transformer certaines des variables pour les rendre davantage exploitables au niveau économétrique. Ainsi, les sections suivantes sont destinées à présenter les modifications que nous avons apportées aux données d'origine pour les conformer aux prérequis de nos analyses à venir.

B. Données manquantes

Les deux bases originelles associées à l'enseignement des mathématiques et du portugais (trois notes, absences, cours particuliers) contiennent respectivement 395 et 649 individus. Du fait de notre intention de recenser les données dans une unique base, nous avons été confrontés au problème suivant : dans la mesure où les deux jeux de données ne comportent pas le même nombre d'observations, nous pouvons en déduire que les informations des étudiants concernant à la fois les mathématiques et le portugais n'ont pas toutes été recensées. Autrement dit, nous ne disposons pas des informations liées aux deux matières pour chacun de nos individus.

Ainsi, pour pallier ce problème et construire notre nouveau jeu de données sur la base d'une ligne par individu recensant à la fois les notes, absences et cours particuliers de mathématiques et de portugais, nous avons dû réussir à retrouver dans les bases les deux lignes associées à un unique individu, puis rejeter les observations pour lesquelles il manquait l'une ou l'autre matière. Pour ce faire, nous nous sommes servis des caractéristiques sociales/émotionnelles/démographiques et autres informations liées à l'éducation (indépendamment de la matière considérée), soit 28 variables, pour associer deux notes à l'individu concerné, en posant l'hypothèse qu'aucune

confusion entre deux individus n'était possible au vu du grand nombre de variables impliquées. Plus techniquement parlant, nous avons concaténé le contenu de toutes les colonnes concernées pour créer dans une nouvelle colonne un identifiant unique qui nous permet d'une part distinguer deux individus différents, et par extension d'identifier les « doublons » pour rattacher à chaque étudiant les informations spécifiques liées aux mathématiques et au portugais. Nous avons pu remarquer que chaque étudiant ayant une note de mathématiques avait nécessairement une note de portugais, tandis que l'inverse n'est pas forcément vérifié dans la mesure où 254 individus n'avaient que des renseignements sur l'enseignement du portugais et n'ont donc pas été conservés dans notre base finale.

Nous avons ainsi obtenu une base constituée de 395 observations et 38 variables issues des caractéristiques socio-démographiques, socio-émotionnelles et concernant l'éducation et les résultats scolaires de chacun de nos individus. Pour rendre nos données utilisables d'un point de vue économétrique, nous avons dû nettoyer les données en modifiant certaines variables. Les différentes transformations appliquées à notre jeu de données sont présentées dans la sous-partie ci-après.

C. Transformations appliquées

Notre jeu de données et plus précisément les modalités des variables qui le composent ont subi diverses transformations pour que nous puissions réaliser des estimations économétriques dessus. Nous tâcherons ici de rappeler de façon non-exhaustive les différentes modifications appliquées aux variables d'origine, puis nous présenterons nos nouvelles variables plus spécifiquement dans la dernière sous-partie.

Dans un premier temps, vingt variables parmi les 38 qui composent le jeu de données sont des variables numériques. Ces dernières n'ont donc pas nécessité d'être retravaillées pour être exploitables. Néanmoins, les dix-huit variables restantes sont des variables catégorielles à deux modalités ou plus, ce qui implique une transformation en variable binaires pour pouvoir être utilisées.

Ainsi, les variables résultant d'une question à réponse fermée "oui"/"non" ont simplement été transformées par une variable binaire en remplaçant le "oui" par 1 et le "non" par 0. Pour d'autres variables créées à partir d'une question à réponse fermée impliquant deux modalités (sexe, établissement scolaire, etc.), nous avons décidé de façon arbitraire du statut de l'individu "de référence" qui correspond à la valeur "0" quand la modalité restante prend la valeur "1". De plus, pour les variables qui présentaient plus de deux modalités, nous avons fait le choix de grouper les modalités de façon cohérente pour ne construire qu'une seule variable binaire en sortie. Enfin, deux de nos variables à plus de deux modalités qui concernent respectivement l'activité professionnelle de la mère et du père ont subi un traitement spécifique : nous avons décidé de regrouper ces deux colonnes pour n'en créer qu'une seule en sortie. Notre jeu de données final comporte ainsi 37 variables, et nous tâcherons d'apporter une courte présentation pour chacune d'elles dans la sous-partie suivante.

D. Présentation des variables

Comme introduit dans la sous-partie précédente, la base de données retravaillée utilisée dans le cadre de notre étude comporte 37 variables dont 20 quantitatives discrètes et 17 quantitatives binaires. Nous les avons regroupés de façon non-exhaustive par catégorie et avons apporté une petite présentation à chacune ci-après.

i. Variables socio-démographiques

woman : prend la valeur 1 si l'individu est une femme, 0 s'il s'agit d'un homme.

age : âge de l'individu (entre 15 et 22 ans).

rural : prend la valeur 1 si l'individu réside dans un milieu rural, 0 s'il réside dans un milieu urbain.

GT3 : prend la valeur 1 si l'individu vit dans un foyer de plus de trois personnes, 0 sinon.

Ptogether : prend la valeur 1 si les parents de l'individu vivent ensemble, 0 sinon.

Medu : représente le niveau d'éducation de la mère, les valeurs étant comprises entre 0 (le plus faible) et 4 (le plus élevé).

Fedu : représente le niveau d'éducation du père, les valeurs étant comprises entre 0 (le plus faible) et 4 (le plus élevé).

Pteacher : prend la valeur 1 si l'individu a au moins un parent enseignant, 0 sinon.

Mguardian : prend la valeur 1 si le tuteur de l'enfant est sa mère, 0 sinon.

ii. Variables socio-émotionnelles

internet : prend la valeur 1 si l'individu a un accès internet à la maison, 0 sinon.

romantic : prend la valeur 1 si l'individu est dans une relation amoureuse, 0 sinon.

famrel : représente la qualité des relations familiales, les valeurs étant comprises entre 1 (très mauvaises) et 5 (excellentes).

freetime : représente le temps libre après l'école, les valeurs étant comprises entre 1 (très peu de temps) et 5 (beaucoup de temps).

goout : représente le fait de sortir avec des amis, les valeurs étant comprises entre 1 (très peu de sorties) et 5 (beaucoup de sorties).

Dalc : représente la consommation d'alcool en semaine, les valeurs étant comprises entre 1 (très peu) et 5 (beaucoup).

Walc : représente la consommation d'alcool pendant le week-end, les valeurs étant comprises entre 1 (très peu) et 5 (beaucoup).

health : représente la condition de santé actuelle, les valeurs étant comprises entre 1 (très mauvaise) et 5 (très bonne).

iii. Variables éducation

GP : prend la valeur 1 si l'individu étudie à l'école Gabriel Pereira, 0 si il étudie à l'école Mousinho da Silveira.

reason : prend la valeur 1 si l'individu a choisi l'école pour sa réputation ou la préférence pour les cours, 0 s'il l'a choisi pour sa proximité de la maison ou pour une autre raison.

traveltime : représente le temps pour faire le trajet maison-école, les valeurs étant comprises entre 1 et 4 (1 - <15 min, 2 - 15 à 30 min, 3 - 30 min à 1 heure, ou 4 - >1 heure).

studytime : représente le temps hebdomadaire consacré aux études, les valeurs étant comprises entre 1 et 4 (1 - <2 heures, 2 - 2 à 5 heures, 3 - 5 à 10 heures, ou 4 - >10 heures).

failures : représente le nombre de fois où l'individu n'a pas validé une matière (4 étant la valeur maximale, même s'il en a plus à son actif).

schoolsup : prend la valeur 1 si l'individu fait du soutien scolaire, 0 sinon.

famsup : prend la valeur 1 si l'individu reçoit du support éducatif de la part de sa famille, 0 sinon.

paid-por : prend la valeur 1 si l'individu a payé pour des cours particuliers en portugais, 0 sinon.

paid-mat : prend la valeur 1 si l'individu a payé pour des cours particuliers en mathématiques, 0 sinon.

activities : prend la valeur 1 si l'individu participe à des activités périscolaires, 0 sinon.

nursery : prend la valeur 1 si l'individu a été à l'école maternelle, 0 sinon.

higher : prend la valeur 1 si l'individu souhaite poursuivre ses études, 0 sinon.

absences-por : nombre d'absences recensées en cours de portugais (entre 0 et 93)

absences-mat : nombre d'absences recensées en cours de mathématiques (entre 0 et 93)

G1-por : note de portugais en première période (de 0 à 20)

G1-mat : note de mathématiques en première période (de 0 à 20)

G2-por : note de portugais en seconde période (de 0 à 20)

G2-mat : note de mathématiques en seconde période (de 0 à 20)

G3-por : note finale de portugais (de 0 à 20)

G3-mat : note finale de mathématiques (de 0 à 20)

Chapitre 1 : Méthodes d'évaluations économiques

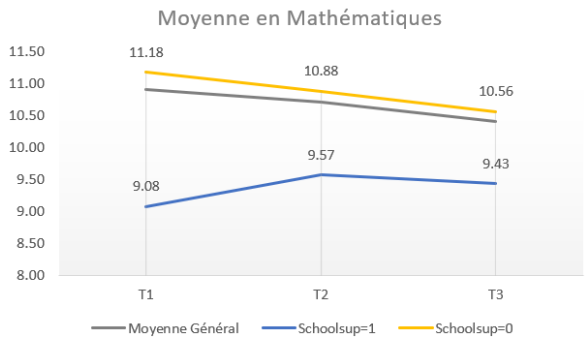
III. Impact du support scolaire sur les moyennes générales des étudiants

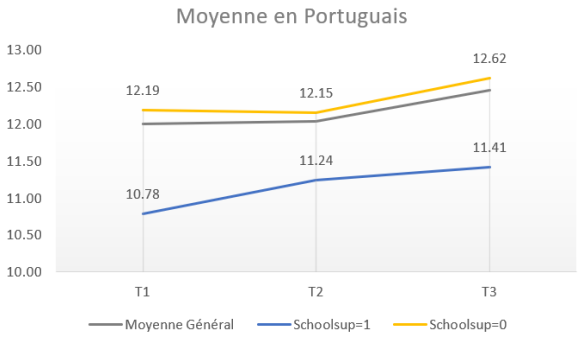
Dans cette partie nous allons utiliser la méthode de l'appariement par score de propension. Dans notre exemple, les personnes traitées seront les personnes ayant eu accès à des cours de support scolaire (autrement dit, la variable schoolsup). Ensuite, les étudiants seront appariés via leurs notes en mathématiques et en portugais, au cours du premier trimestre. Nous allons opérer différentes variantes de cet appariement. L'objectif de cette méthode est de montrer que même les étudiants ayant les plus mauvaises notes, s'ils ont bénéficié de support scolaire, ont plus progressé que les étudiants ayant des meilleures notes mais n'ayant pas bénéficié de ce dernier.

A. Statistiques Descriptives sur l'impact du support scolaire

i. Schoolsup

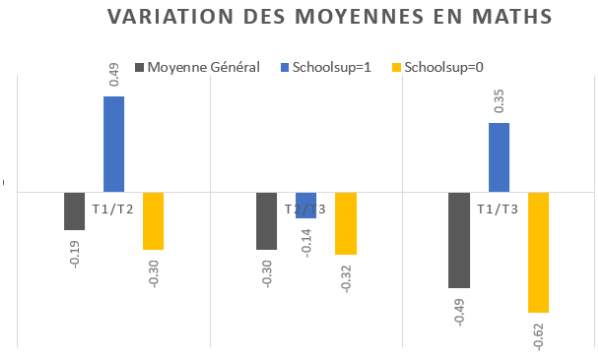
Pour commencer, nous allons comparer les moyennes des élèves ayant bénéficié de cours particuliers, avec la moyenne des élèves n'en ayant pas bénéficié et la moyenne générale. Les élèves ayant bénéficié du support scolaire sont au nombre de 51, contre 344 n'en ayant pas profité. Nous aurons d'une part les moyennes en mathématiques, et de l'autre les moyennes en Portugais.

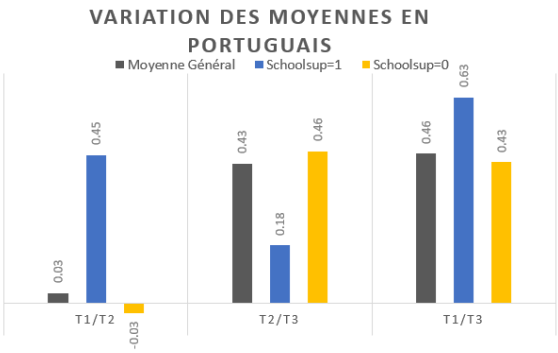




Les résultats sont similaires sur les 2 graphiques, c'est-à-dire, que les étudiants ayant bénéficié du support scolaire se place bien en dessous de la moyenne générale contrairement à ceux n'en ayant pas bénéficié se trouvant au-dessus. Cela semble donc montrer que ce sont les étudiants les plus en difficultés qui bénéficient de ce dispositif.

Cependant si l'on s'intéresse à la progression des étudiants, c'est à dire, à la variation de moyenne entre 2 instants dans le temps (ici nous avons réalisé 3 comparaisons, une entre le T1 et le T2, entre le T2 et le T3, pour finir, entre le T1 et le T3).

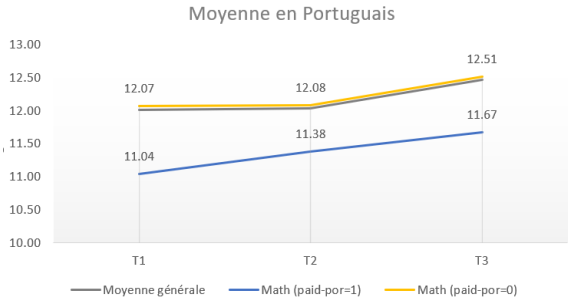


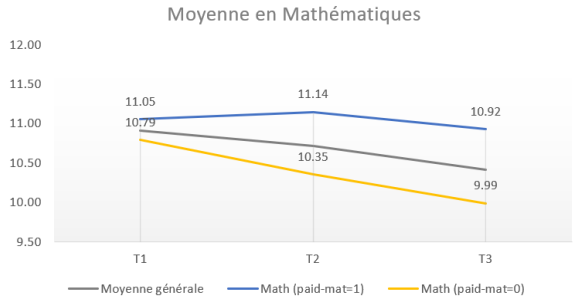


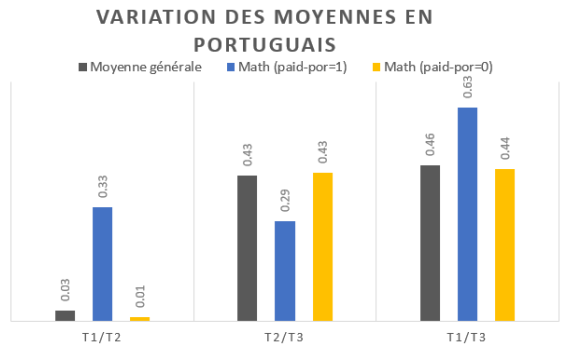
Nous pouvons voir qu'il y a pour les moyennes en mathématiques, une plus grande progression (ou une moindre régression) pour le groupe des étudiants ayant bénéficié du support scolaire, avec une progression de 0,35 sur l'ensemble de l'année contre une baisse de 0.49 de la moyenne générale. Le même phénomène est observable pour le Portugais, sauf pour la différence entre le T2 et le T3, ou les étudiants ayant bénéficié du support scolaire ont moins progressé que la moyenne. Cependant, si l'on observe la progression sur l'ensemble de l'année, les élèves ayant bénéficié du support scolaire ont gagné en moyenne 0.63 points sur leurs moyennes, contre 0.46 en générale. Il semble donc que le soutien scolaire à un impact positif sur les étudiants, même s'il reste faible.

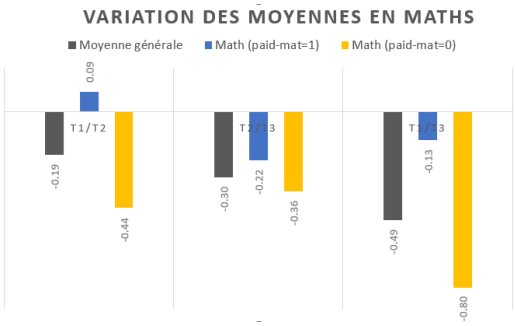
ii. Paid-por & Paid-mat

Nous avons ensuite décidé d'analyser l'impact du support scolaire en mathématiques et en portugais (autrement dit les variables "paid-por" et "paid-mat"). Ce support scolaire est légèrement différent car d'une part, il est spécialisé dans une matière particulière. D'autre part, ce support scolaire est au frais des parents de l'étudiant. Les étudiants ayant bénéficié du support scolaire en mathématiques sont 181 contre 24 pour les cours particuliers en portugais.







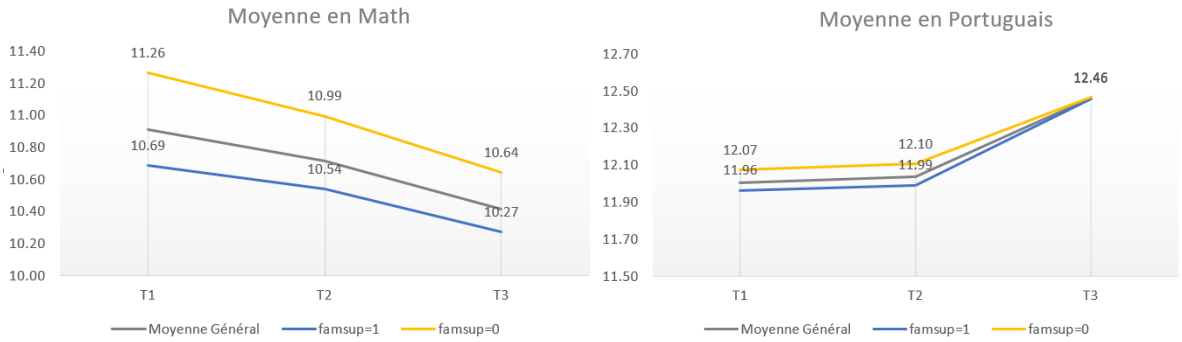


Sur ces graphiques nous pouvons donc constater 2 phénomènes bien différents, avec d'une part le soutien scolaire en mathématiques, dont près de la moitié des étudiants ont recours. Ces étudiants ont un niveau supérieur à la moyenne en mathématiques et sont ceux dont les moyennes diminuent le moins au cours de l'année (baisse de 0,13 points en moyenne contre -0.49 en générale et -0.80 pour les étudiants n'ayant pas eu de support scolaire).

A l'inverse, les personnes ayant suivi du support scolaire en Portugais sont ceux qui ont de base un niveau bien inférieur à la moyenne (11.04 contre 12.07 en moyenne). Cependant, ces étudiants sont également ceux qui ont le plus progressé en moyenne avec une hausse de 0.63 points sur leurs moyennes contre 0.46 en générale.

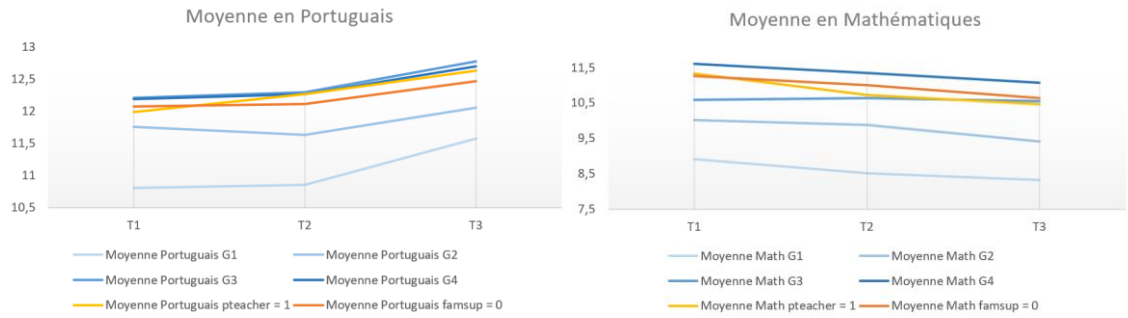
iii. Famsup

Pour finir sur les variables liées au support scolaire, nous allons étudier la variable famsup qui correspond au support éducatif fournis par la famille. Il y a 242 étudiants qui en ont bénéficié (contre 153 qui n'en ont pas bénéficié), c'est la forme de support scolaire dont les étudiants ont le plus bénéficié. Potentiellement, cette forme de support est également la moins fiable car elle n'est peut-être pas aussi régulière pour tous les étudiants. De plus, elle n'est pas dispensée avec la même qualité par l'ensemble des parents, ce sont des éléments que nous prendrons en compte par la suite.



Comme nous pouvons le voir, les résultats ne sont pas "logiques" ou alors ils ne sont pas vraiment significatifs, par exemple, les moyennes en portugais sont très proches et finissent à la même moyenne.

Cependant, si nous couplons la variable famsup avec les variables Medu et Fedu, qui représentent le niveau d'étude des parents, allant de 1 à 4. Pour simplifier l'analyse de cette variable, nous avons créé une nouvelle variable calculant la moyenne d'éducation des 2 parents, que nous arrondissons à l'unité, afin de conserver les 4 niveaux d'étude. De plus, nous avons utilisé la variable "Pteacher", codé 1, si l'un des parents est professeur.



Sur le graphique ci-dessus, nous retrouvons les 4 groupes (qui sont la moyenne de niveau d'études des parents (4 étant le plus élevé)) en bleu. Ensuite, en rouge, la moyenne des personnes n'ayant pas bénéficié de soutien de la part de leurs parents, et pour finir, la moyenne des personnes ayant un parent professeur en jaune.

Nous pouvons donc dire qu'il semble exister une relation linéaire entre le niveau d'étude des parents et la variable famsup. En effet, plus le niveau d'étude des parents est élevé, plus l'efficacité du support scolaire sera élevée, ce qui amènera à des moyennes de plus en plus élevées. Concrètement, on remarque que les étudiants dont les parents ont un niveau d'étude compris

entre 1 et 2 ont des moyennes inférieures aux étudiants n'ayant pas bénéficié du support scolaire familiale, contrairement à ceux ayant des parents ayant un niveau d'étude compris entre 3 et 4. Les étudiants ayant au moins un parent professeur et bénéficiant du support scolaire ont une moyenne s'approchant des étudiants n'ayant pas bénéficié du support scolaire.

Cependant, si l'on analyse les moyennes des étudiants, indépendamment de leurs accès aux soutien scolaire de la part de la famille, nous retrouvons le même phénomène. Il semble donc que nous ne pouvons rien conclure sur la variable famsup pour le moment.

B. Méthode d'appariement par score de propension

i. Explication de la méthode

Pour étudier plus en détail l'impact du support scolaire, nous allons utiliser la méthode de l'appariement par score de propension. L'objectif de cette courte introduction est donc de présenter le fonctionnement de la méthode et d'expliquer comment nous allons procéder.

Ainsi, l'appariement (ou matching) est une méthode servant à estimer les effets causaux d'un programme d'intervention, il y aura des individus traités et des individus non traités. De plus, la méthode cherche à associer les individus traités avec 1 ou plusieurs individus non traités selon des caractéristiques proches (autrement dit, de créer un contrefactuel)[1]. Dans cette partie, nous allons apparier les individus ayant reçu du support scolaire (variable schoolsup) avec ceux n'en ayant pas reçu. Les étudiants seront couplés en fonction de leurs notes au premier trimestre, l'objectif étant d'étudier la variation des notes entre le premier trimestre et le dernier. Nous cherchons à prouver l'hypothèse selon laquelle les étudiants ayant bénéficié du support scolaire ont plus progressé que les étudiants d'un même niveau mais n'en n'ayant pas bénéficié. Puis nous allons répéter cette méthode afin d'étudier l'impact des autres formes de support scolaire. Au total, il y a 4 formes de support scolaire, le schoolsup, venant de l'école, le paid-mat/paid-por, qui sont des cours payants et le famsup, qui correspond au support scolaire venant de la famille.

ii. Variable schoolsup

Pour cette première variable, nous avons donc séparer l'échantillon en 2 groupes (schoolsup=1/schoolsup=0), puis nous estimer un score de propension à partir des notes du premier trimestre. L'objectif était d'appareiller les étudiants par groupe de niveau car comme nous l'avons vu, les étudiants ayant bénéficié du support scolaire ont des notes inférieures à la moyenne. Puis nous avons estimer la variation de moyenne entre le premier trimestre et le second.

a. Impact sur la moyenne en mathématique

Pour commencer nous allons estimer l'impact sur la moyenne en mathématique.

Ainsi nous obtenons le code suivant :

[1] Chiappori, Pierre-André. « Modèles d'appariement en économie. Quelques avancées récentes ». *Revue économique*, vol. 63, nº 3, 2012, p. 437-52. *Cairn.info*, https://doi.org/10.3917/reco.633.0437.

```
Y <- mybase$progmat # variable de progression (T3-T1)

Tr <- mybase$schoolsup # variable de traitement (support scolaire)

Match.out <- Match(Y = Y, Tr = Tr, estimand="ATT",M=2, replace = TRUE, ties = TRUE, caliper
= 0.15, BiasAdjust = TRUE, X = cbind(mybase$G1mat), exact = c("TRUE"))
```

Résultat obtenue :

ESTIMATION AJUSTE	COEFFICIENT	ESTIMATION NON AJUSTE	COEFFICIENT
ESTIMATES	1.3486	ESTIMATES NOADJ	1.3486
AI SE	0.382	SE	0.56513
T-STATS	3.5304	T-STATS	2.3864
P-VALUE	0.00041487	P-VALUE	0.017013
Treated obs = 51 Matched obs = 51 $\bar{x} = -0.49$			

Ainsi, nous pouvons donc dire que l'effet moyen des cours particuliers fournis par l'école est une augmentation de 1.34 point de moyenne en mathématique entre le premier et le troisième trimestre, alors que la moyenne chute de 0.49 points au cours de l'année (en moyenne). Il y a donc un effet élevé et positif du support scolaire fournis par l'école. De plus, nous avons un R^2 significatif au seuil de 5% ce qui viens confirmer nos résultats.

b. Impact sur la moyenne en portugais

Désormais, nous allons nous intéresser à l'impact du dispositif sur la progression en portugais.

Ainsi nous obtenons le code suivant :

```
Y <- mybase$progpor # variable de progression (T3-T1)

Tr <- mybase$schoolsup # variable de traitement (support scolaire)

Match.out <- Match(Y = Y, Tr = Tr, estimand="ATT",M=2, replace = TRUE, ties = TRUE, caliper
= 0.15, BiasAdjust = TRUE, X = cbind(mybase$G1por), exact = c("FALSE"))
```

Résultat obtenue :

ESTIMATION AJUSTE	COEFFICIENT	ESTIMATION NON AJUSTE	COEFFICIENT
ESTIMATES	-0.13811	ESTIMATES NOADJ	-0.13811
AI SE	0.25026	SE	0.3107
T-STATS	-0.55187	T-STATS	-0.44452
P-VALUE	0.58104	P-VALUE	0.65667
Treated obs = 51 Matched obs = 50 $\bar{x} = 0.46$			

Ainsi, il semble qu'avec cette nouvelle estimation, nous obtenons les résultats inverse, avec une baisse de -0.138 point de moyenne en portugais lorsque les étudiants participent au soutien scolaire. Alors qu'en moyenne les étudiants gagne 0.46 point au cours de l'année scolaire dans cette matière). Cependant, ces résultats sont à relativiser, en effet, en plus d'être très faible, ils ne sont pas significatifs. Cela s'explique peut-être par la nature de ce support scolaire, en effet, nous ne connaissons pas la matière/les matières qui sont concernées par le support scolaire. Donc nous ne pouvons rien conclure sur l'impact du soutien scolaire sur les notes en portugais.

iii. Variable paid-por & paid mat

Lors de cette seconde partie, nous allons estimer l'impact des variables paid-por et paid-mat les moyennes des étudiants. Ce sont les cours qui selon nous, sont les plus efficaces pour faire augmenter les notes des étudiants.

a. Impact sur la moyenne en mathématique

Pour commencer nous allons estimer l'impact sur la moyenne en mathématique.

Ainsi nous obtenons le code suivant :

```
Y <- mybase$progmat # variable de progression (T3-T1)

Tr <- mybase$paidmat # variable de traitement (support scolaire)

Match.out <- Match(Y = Y, Tr = Tr, X = cbind(mybase$G1mat), estimand = "ATT", exact =
c("TRUE","FALSE","FALSE"), M=2, caliper = 0.15, sample=TRUE)
```

Résultat obtenue :

ESTIMATION AJUSTE	COEFFICIENT	ESTIMATION NON AJUSTE	COEFFICIENT
ESTIMATES	0.66922	ESTIMATES NOADJ	0.66922
AI SE	0.28934	SE	0.28611
T-STATS	2.313	T-STATS	2.339
P-VALUE	0.020724	P-VALUE	0.019335
Treated obs = 181 Matched obs = 181 $\bar{x} = -0.49$			

Ainsi, nous pouvons constater que l'effet moyen du soutien scolaire payant fait augmenter de 0.66 point la moyenne en mathématique entre le premier et le troisième trimestre, alors que la moyenne chute de 0.49 points au cours de l'année (en moyenne). Cela reste cependant moins efficace que le support scolaire fournis par l'école. Cela peut potentiellement s'expliquer par le fait que les étudiants qui paie pour des cours en math ont déjà des notes supérieur à la moyenne, la progression est donc plus difficile, que pour les autres étudiants ayant des notes biens inférieur à la moyenne.

b. Impact sur la moyenne en portugais

Pour cette seconde partie, nous allons analyser l'impact des cours payant en portuguais.

Ainsi nous obtenons le code suivant :

```
Y <- mybase$progpor # variable de progression (T3-T1)

Tr <- mybase$paidpor # variable de traitement (support scolaire)

Match.out <- Match(Y = Y, Tr = Tr, X = cbind(mybase$G1por), estimand = "ATT", exact =
c("FALSE","FALSE","FALSE"), M=2, caliper = 0.15, sample=TRUE)
```

Résultat obtenue :

ESTIMATION AJUSTE	COEFFICIENT	ESTIMATION NON AJUSTE	COEFFICIENT
ESTIMATES	0.86205	ESTIMATES NOADJ	0.86205
AI SE	0.56523	SE	0.76251
T-STATS	1.5251	T-STATS	1.1305
P-VALUE	0.12723	P-VALUE	0.25825

Treated obs = 24 Matched obs = 24 $\bar{x} = 0.46$

Ces résultats sont plus encourageants, car ce support scolaire fais augmenter de 0.86 point la moyenne des étudiants par rapport à ceux qui n'ent ont pas bénéficier. Cependant, les résultats ne sont pas significatifs. Selon nous, cela est lié a la faible quantité d'observation traité (24).

iv. Variable famsup

Lors de cette troisième et dernière partie, nous allons estimer l'impact de la variable famsup sur les moyennes des étudiants. La légère originalité ici est que l'on calcul le score de propension à partir des variable «Medu » et « Fedu » en plus des notes du premier trimestre.

a. Impact sur la moyenne en mathématique

Pour commencer nous allons estimer l'impact sur la moyenne en mathématique.

Ainsi nous obtenons le code suivant :

```
Y <- mybase$progmat # variable de progression (T3-T1)

Tr <- mybase$famsup # variable de traitement (support scolaire)

Match.out <- Match(Y = Y, Tr = Tr, X = cbind(mybase$G1mat,mybase$Medu,mybase$Fedu), estimand
= "ATT", exact = c("TRUE","FALSE","FALSE"), M=2, caliper = 0.15, sample=TRUE)
```

Résultat obtenue :

ESTIMATION AJUSTE	COEFFICIENT	ESTIMATION NON AJUSTE	COEFFICIENT
ESTIMATES	-0.33825	ESTIMATES NOADJ	1.3486
AI SE	0.19346	SE	0.46504
T-STATS	-1.7485	T-STATS	-0.72736
P-VALUE	0.080383	P-VALUE	0.46701
Treated obs = 242 Matched obs = 61 $\bar{x} = -0.49$			

Ainsi, nous pouvons donc dire que l'effet moyen du soutien scolaire familiale font diminuer de -0.33 point de moyenne en mathématique entre le premier et le troisième trimestre, alors que la moyenne chute de 0.49 points au cours de l'année (en moyenne). Cela implique donc que les notes en maths diminuent, mais diminue légèrement moins qu'en par rapport à la moyenne. Nous avons réestimer en retirant Medu et Fedu de l'appariement et le coefficient deviens positif mais non significatif (Coef : 0.34311 / p-value : 0.2667). Ainsi nous pouvons conclure que le soutien scolaire familiale semble avoir un impact légèrement positif sur la moyenne scolaire des étudiants.

b. Impact sur la moyenne en portugais

Nous obtenons le code suivant :

```
Y <- mybase$progpor # variable de progression (T3-T1)

Tr <- mybase$famsup # variable de traitement (support scolaire)

Match.out <- Match(Y = Y, Tr = Tr, X = cbind(mybase$G1por,mybase$Medu,mybase$Fedu), estimand
= "ATT", exact = c("FALSE","FALSE","FALSE"), M=2, caliper = 0.15, sample=TRUE)
```

Résultat obtenue :

ESTIMATION AJUSTE	COEFFICIENT	ESTIMATION NON AJUSTE	COEFFICIENT
ESTIMATES	0.22416	ESTIMATES NOADJ	1.3486
AI SE	0.099442	SE	0.15151
T-STATS	2.2542	T-STATS	1.4795
P-VALUE	0.024185	P-VALUE	0.139
Treated obs = 242 Matched obs = 109 $\bar{x} = 0.46$			

Ces résultats sont plus encourageants, car en plus d'être significatif au seuil de 5%. Cependant, même si ces résultats semblent plus encourageants (car positif), l'impact du support familial est plus faible que la progression moyenne dans la matière. Ainsi, nous pouvons donc dire que l'effet moyen du soutien scolaire familiale fait augmenter de 0.22 point la moyenne en portugais entre le premier et le troisième trimestre, alors que les étudiants progressent en moyenne de 0.46 points au cours de l'année. Cela va ainsi dans le même sens inverse que pour la moyenne en mathématiques, même si les effets restes minime.

v. Récapitulatif des résultats

Pour finir, nous avons décider de réaliser un tableau récapitulatif de l'ensemble de nos résultats :

TRAITEMENT	VARIATION EN MATHEMATIQUE	VARIATION EN PORTUGAIS
SCHOOLSUP	1.3486***	-0.13811
PAID-POR & PAID MAT	0.66922**	0.86205
FAMSUP	-0.33825*	0.22416**

Ainsi, nous pouvons remarquer que le dispositif le plus efficace pour faire progresser les étudiants sont le cours offert par l'école ou les cours payants. A l'inverse, pour l'espagnol il n'y a pas vraiment de dispositif permettant de faire augmenter la moyenne des étudiants. Il aurais fallu avoir une plus grande part d'étudiant suivant des cours de portugais.

IV. Impact de la volonté de poursuivre des études supérieures sur la réussite scolaire des étudiants.

Lors de cette seconde partie, nous avons pour objectif d'étudier l'impact de la volonté propre des étudiants sur leurs réussites scolaires. Plus précisément, nous voulons montrer qu'un étudiant ayant la volonté de réussir ses études scolaires aura un impact positif sur la moyenne. L'objectif de cette partie est de montrer que lorsque qu'un étudiant à la volonté de poursuivre des études scolaire, l'impact des comportements "démotivant" sera plus faible que pour un étudiant n'ayant pas cette volonté. Dans ce cadre, nous nous servirons de plusieurs variables qui sont signe de motivation, comme "higher" codé 1 si la personne souhaite poursuivre des études supérieures.

A. Hypothèses

Dans cette première partie, nous avons récapitulé les effets attendus des variables que nous allons utiliser sur les moyennes des étudiants.

VARIABLES	HYPOTHESE (HIGHER=1)	HYPOTHESE (HIGHER=0)
HIGHER	+	-
PAID MAT & PAID POR	+	-
INTERNET	+ +	+
ROMANTIC	+	+ =
FREETIME	+	=
GOOUT	+	-
DALC & WALC	-	- -

B. Statistiques descriptives

Dans le cadre de la seconde partie, nous allons réaliser une régression multiple en divisant l'estimation avec la variable "higher". L'objectif est d'estimer l'impact de comportement n'encourageant pas la réussite scolaire comme par exemple, la consommation d'alcool où le nombre de sorties.

Dans un premier temps, nous faisons donc l'hypothèse que ces comportements ont un effet négatif sur la réussite scolaire des étudiants. Dans un second temps, nous faisons l'hypothèse que ces comportements ont un impact plus faible sur les étudiants n'ayant qu'une moindre velléité de suivre leurs études.

Pour estimer cette volonté de suivre ses études, ou au moins de réussite scolaire, nous avons utilisé la variable higher qui est la plus appropriée. Cependant il n'y a que 20 étudiants ne souhaitant pas poursuivre leurs études ce qui représente un échantillon trop faible, nos résultats risquent d'être biaisés (aussi appelé biais d'échantillonnage). Ainsi nous avons choisi une seconde variable, la variable paid-mat et ce pour plusieurs raisons.

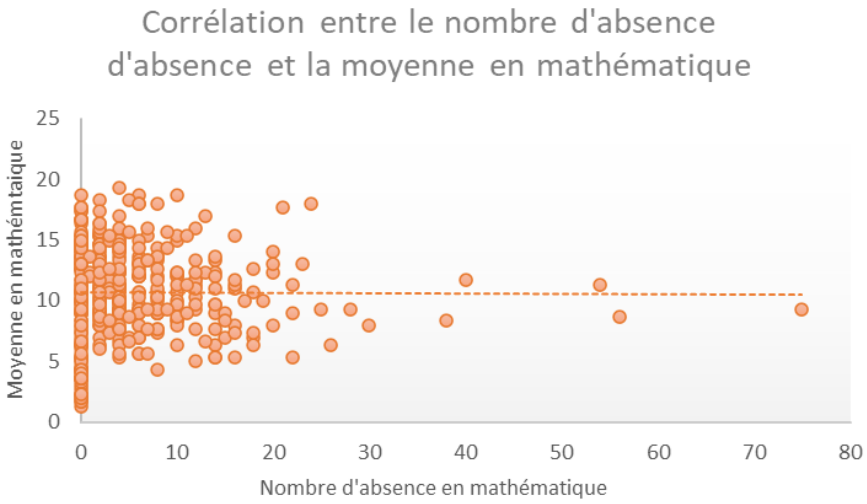
Dans un premier temps, cette variable est représentative d'une certaine volonté de l'étudiant (ou des parents de l'étudiant) de réussir ses études car il ira payer ces cours particuliers. Dans un second temps, cette variable n'est pas soumise au biais de taille d'échantillonnage (181 étudiants en ont bénéficié).

Cependant, il existe tout de même un biais, en effet, comme nous ne connaissons pas le prix de ce cours particulier en mathématiques, il se peut que certains étudiants n'aient pas participé pour des raisons économiques. Ainsi, cela pourrait exclure certains étudiants qui pourraient avoir l'envie de participer à ces cours mais qui n'en ont pas les moyens.

i. Effet du nombre d'abence

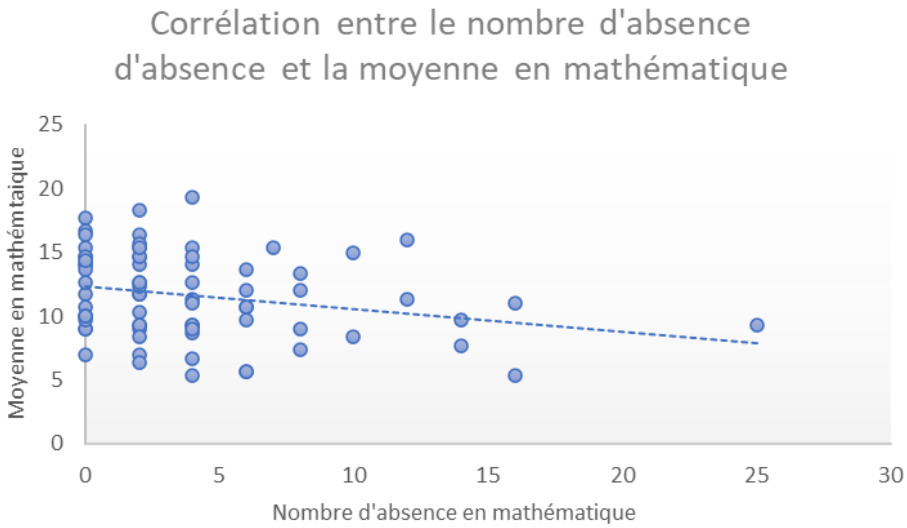
Dans cette partie nous avons estimé les effets de 3 variables sur la moyenne. Ces variables sont le nombre de jours d'absence, la consommation d'alcool et la quantité de sortie.

Ainsi, ci-dessous, nous avons réalisé un nuage de point croisant d'une part, la moyenne de l'étudiant en math et le nombre d'absence dans cette matière. De plus, nous avons rajouté une courbe de tendance qui représente la nature de la relation entre les 2 variables.



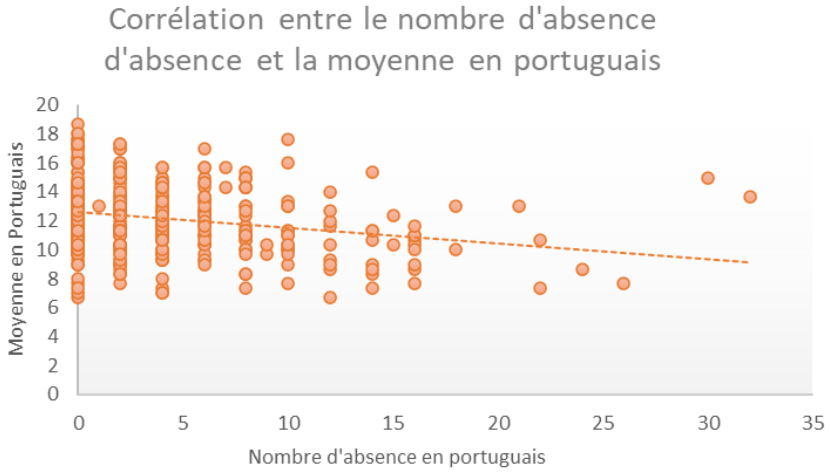
Comme nous pouvons ainsi constater qu'il ne semble pas avoir de lien entre la moyenne et les absences. Cependant, il n'y a que 5 étudiants qui ont eu plus de 30 jours de retard.

Ainsi, comme le nombre de jours d'absences est parfois "atypique", nous avons décidé de sélectionner seulement les étudiants dont le nombre d'absence est inférieur à 30. Cette transformation permet donc d'obtenir le nuage de point ci-dessous :



Dans le graphique ci-dessus, la tendance est plus claire, avec une légère baisse de la moyenne lorsque le nombre d'absence augmente. Ces résultats sont logique, même si les effets des absences semble n'avoir qu'un faible impact sur la note des étudiants.

Nous avons réalisé le même graphique ci-dessous, représentant la corrélation entre la moyenne en portuguais et les absences dans cette même matière.

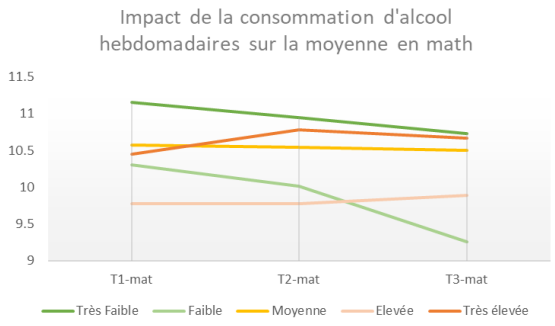


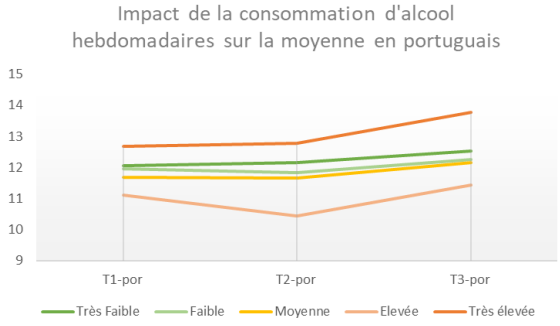
Comme nous pouvons le constater, les moyennes diminuent lorsque le nombre d'absence augmente, ce qui est un phénomène logique. Encore une fois, l'influence des absences sur la moyenne reste faible.

ii. Effet de la consommation d'alcool

Dans cette seconde partie nous allons analyser l'impact de la consommation d'alcool. Pour cela, nous avons divisé l'échantillon en 5 groupes, pour chaque groupe de consommateur d'alcool (1 = Très faible consommation et 5 = consommation très élevée). Puis nous avons calculé la moyenne à chaque trimestre pour chaque groupe et pour chaque matière. De plus, nous avons estimé d'une part, l'impact de la consommation d'alcool en semaine et d'autre part, l'impact de la consommation durant le weekend.

Les 2 graphiques ci-dessous montrent donc la moyenne en math et en portugais de chaque catégorie. Plus la courbe est verte, moins la personne boit de l'alcool et inversement pour le rouge.

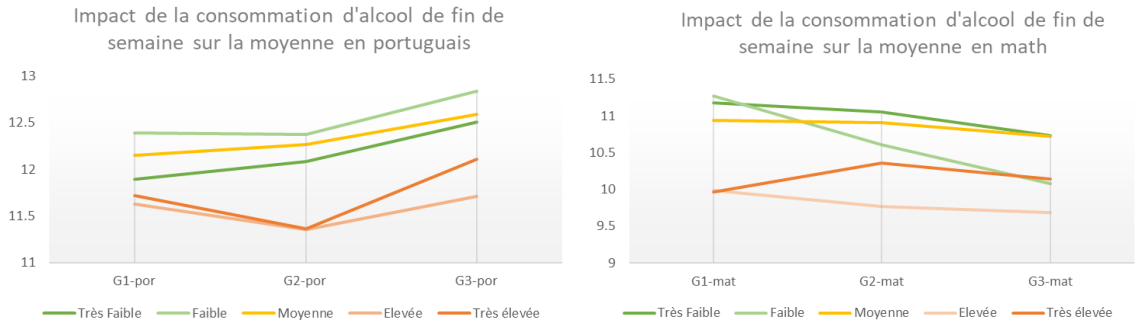




Dans un premier temps, nous pouvons constater que la relation n'est absolument pas linéaire, avec des étudiants qui boivent en faible quantité qui finissent dernier. Plus étonnant encore, en portugais, les étudiants qui ont les meilleures moyennes sont ceux qui boivent le plus, sans cette composante, les résultats sont relativement cohérents. Ce qui explique ces résultats "singuliers" est la taille de chaque groupe et de l'échantillon. En effet, il n'y a que peu d'étudiants qui

s'alcoolisent en semaine, seuls 9 d'entre eux consomment énormément d'alcool, 9 beaucoup et 26 en consomment des quantités "moyennes". Si l'on compare cet échantillon avec les consommateurs d'alcool le week-end, il est moins probable que l'on retrouve ce même biais lié à une erreur d'échantillonnage. En effet, ils sont 28 à en consommer très beaucoup, 51 à en consommer beaucoup et 80 en consomment des quantités "moyenne"

Ainsi, le graphique ci-dessous représente les moyennes scolaires en fonction de la consommation d'alcool durant le week-end.

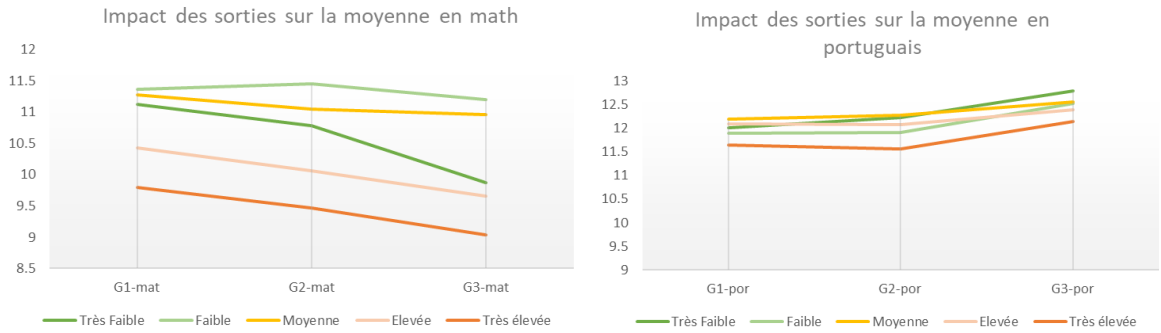


Comme nous pouvons le constater, les résultats sont plus "cohérents" que sur les graphiques précédents. En effet, en portugais, les personnes qui consomment beaucoup d'alcools se détachent de la moyenne et sont inférieures qui en boivent peu, ou faiblement.
On retrouve le même phénomène en mathématique.
Nous pouvons donc conclure que la consommation d'alcool le week-end semble avoir un impact négatif sur la moyenne scolaire. Pour finir, si l'on regroupe les consommateurs "Très faible, faible et Moyen" et les consommateurs "Élevée" et "Très Élevé", et que l'on calcul la moyenne, l'écart est de 0.85 point en math et de 0.7 point en portugais.

iii. Effet des sorties, d'internet, des relations amoureuses et du temps libre

Lors de cette dernière partie, nous nous intéressons à l'effet des autres variables que nous comptons examiner au cours de notre régression multiple.

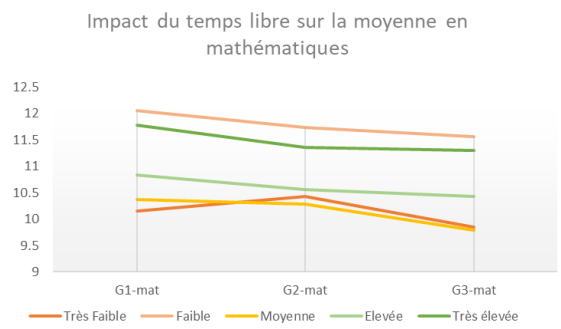
Les graphiques ci-dessous représentent l'impact des sorties sur les moyennes en math et en portugais :

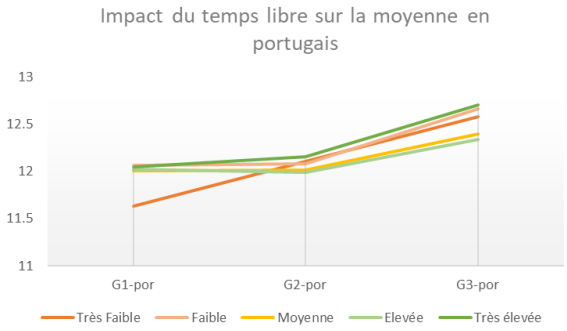


Les résultats sont assez curieux car la quantité de sortie fait varier la moyenne en math beaucoup plus qu'en portugais. De plus, lorsqu'on regarde les résultats en math, il semble que plus la personne fait des sorties, plus la moyenne diminue. Cependant les étudiants qui sortent le moins

se retrouvent au milieu, ces étudiants sont au nombre de 23 et sont peut-être isolés. Nous pouvons donc faire l'hypothèse qu'avoir un cercle social est bénéfique pour un étudiant, lui permettant par exemple de s'entraider en mathématiques, ce qui expliquerait que les étudiants qui sortent qu'un peu on de meilleure moyenne que les autres.

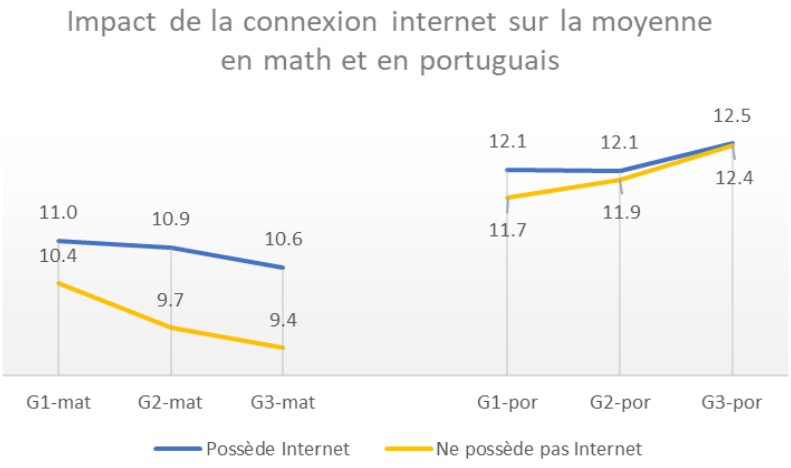
Nous retrouvons sur le graphique ci-dessous l'impact du temps libre sur les moyennes en mathématique et en portugais.





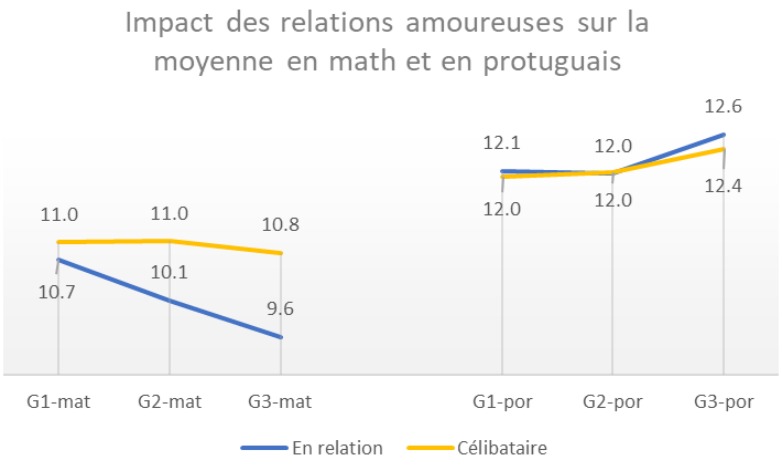
Nous pouvons ainsi constater que l'on retrouve le même phénomène que pour la variable précédente, c'est-à-dire, que les notes en mathématiques varient beaucoup plus que les notes en portugais. Dans un second temps, nous remarquons qu'il ne semble pas y avoir de lien quelconque entre temps libre et moyenne, nous verrons lors de la prochaine partie si cela est vérifié.

Concernant les graphiques ci-dessous, il s'intéresse à l'impact d'internet sur les moyennes des étudiants. Au total, ils sont seulement 66 à ne pas avoir de connexion internet. Comme l'étude date de 2008, ces résultats ne sont pas si étonnant même si ne pas avoir internet reste un marqueur de précarité.



Ainsi, nous pouvons constater que les étudiants n'ayant pas accès à internet ont des moyennes inférieures, cette différence est d'autant plus marquée en mathématique avec près d'un point d'écart entre les 2 groupes. Lorsque l'on s'intéresse à la moyenne en portugais, on remarque que la différence est résiduelle.

Pour finir, le graphique ci-dessous étudie l'impact des relations amoureuses sur les moyennes scolaires entre les étudiants.



Nous remarquons le même phénomène que précédemment, c'est à dire que l'appartenance à l'un des groupes fait varier la moyenne en mathématique mais pas celle de Portugais. De plus, c'est la première fois que nous trouvons des résultats contraires à nos attentes, en effet nous estimons qu'être dans une relation amoureuse faisait augmenter la moyenne alors qu'il semble que c'est l'inverse.

C. Régression multiple

Nous allons désormais utiliser la régression multiple afin de mieux comprendre ce qui quelle est la nature des relations des variables influençant la réussite scolaire.

i. Présentation de la méthode

Dans un premier temps nous chercherons à comprendre si la volonté de poursuite d'études joue réellement un rôle dans la scolarité des élèves, puis nous verrons si le fait que les parents paient des cours du soir impacte la volonté des étudiants à réussir.

Pour ce faire, nous allons réaliser plusieurs régressions multiples selon les différents groupes (ceux qui veulent/ne veulent pas faire d'études supérieurs et ceux qui bénéficient/ne bénéficient pas de cours du soir). Effectivement, il existe des différences entre les étudiants tel que leur capacité à travailler que nous ne pouvons pas contrôler et qui est difficile à contrôler. L'intérêt de créer ces sous-populations nous permet d'éviter une distribution tronquée et de mieux appréhender leurs résultats scolaires en fonction de certaines variables.

Nous allons nous intéresser à la motivation des étudiants et à l'impact de celle-ci sur leurs résultats. Pour ce faire, nous allons donc utiliser les variables suivantes :

- Higher
- Paid-mat (pour les régressions en fonction des notes de math)
- Paid-por (pour les régressions en fonction des notes de portugais)
- Internet
- Romantic
- Freetime

- Goout
- Dalc
- Walc

Pour simplifier l'analyse, les notes des trimestres ont été transformées en moyenne, laissant alors une seule variable : moyenne_mat ou moyenne_por selon les matières.

ii. Régression selon la volonté de poursuivre des études

Nous allons essayer de comprendre si la volonté de faire des études supérieures ou non, influe sur la motivation des élèves et par conséquent leur moyenne. Pour avoir une idée de la répartition de nos élèves, nous pouvons observer ci-contre la répartition des individus en fonction de la volonté à poursuivre les études ou non. On remarque que ceux envisageant des études (en orange) sont surreprésentés car il représente plus de 90 % de notre échantillon tandis que les autres seulement 5 %. Ce sont des informations à prendre en compte dans l'analyse de nos résultats.

C. Estimation de la moyenne en mathématique

Ci-dessous, nous retrouvons un tableau permettant de regrouper trois régressions. La première colonne ne prend pas en compte la différenciation de groupe, tandis que les deux autres représentent les résultats pour les différents groupes.

	TOUT L'ECHANTILLON	VOLONTE DE POURSUIVRE LES ETUDES (HIGHER = 1)	PAS DE VOLONTE DE POURSUIVRE LES ETUDES (HIGHER = 0)
HIGHER	2.753	/	/
PAID-MAT	0.372	0.303	5.97
INTERNET	1.091	1.266	-2.195
ROMANTIC	-0.724	-0.632	-2.497
FREETIME	0.243	0.245	-0.532
GOOUT	-0.558	-0.513	-1.551
DALC	-0.181	-0.158	0.271
WALC	0.036	-0.006	0.725

En moyenne, un étudiant du groupe higher =1 obtient une note de 10.84 tandis qu'un élève ne souhaitant pas poursuivre ses études obtient une note moyenne de 7.65. La création des groupes nous permet de constater qu'une simple régression linéaire ne permet pas de rendre compte réellement de l'ensemble de la population. En effet, si l'on regarde juste la première colonne, on ne réalise pas l'impact réel des cours de mathématique sur la note des élèves : le fait d'avoir des cours du soir, pour ceux qui n'ont pas de réelle volonté de poursuivre les études impact énormément et positivement leur note moyenne. Une grande différence réside dans l'usage

d'internet, pour ceux voulant faire des études cette variable impact positivement sur les notes tandis que pour l'autre groupe cela influe négativement. On peut alors supposer que ce groupe utilise internet moins facilement pour des fins scolaires. Il en va de même pour le temps libre. Étonnement, il semblerait que la consommation d'alcool et de cannabis améliorerait les compétences des étudiants n'ayant pas de volonté de poursuite d'études, cependant cela doit provenir du faible nombre d'individus appartenant à ce groupe (20 contre 375). Avec cette analyse, nous remarquons donc que la motivation d'études supérieures aide les étudiants à rester concentrer dans les études, et à être moins « influencer » par certains comportements (drogues, alcool, sorties…)

d. Estimation de la moyenne en portugais

A présent, intéressons-nous aux notes de portugais.

	TOUT L'ECHANTILLON	VOLONTE DE POURSUIVRE LES ETUDES (HIGHER = 1)	PAS DE VOLONTE DE POURSUIVRE LES ETUDES (HIGHER = 0)
HIGHER	1.022	/	/
PAID-POR	-0.931	-1.039	3.039
INTERNET	0.206	0.213	0.499
ROMANTIC	0.112	0.088	-0.659
FREETIME	0.047	0.085	-1.043
GOOUT	-0.038	-0.093	1.432
DALC	0.065	0.029	0.914
WALC	-0.142	-0.098	-1.673

En moyenne, un étudiant souhaitant poursuivre ses études obtient une note moyenne de 12.22 en portugais tandis que les autres obtiennent une note de 11.17. A l'inverse des notes de maths, le fait d'avoir des cours du soir en portugais impact négativement les notes. Tandis que pour les autres cela leur permet de gagner en moyenne 3 points de plus. On retrouve les mêmes résultats que précédemment pour la variable goout. La consommation d'alcool reste aussi surprenante mais il ne faut pas oublier le faible nombre d'individus dans l'échantillon.

En règle générale, les individus ayant la volonté de faire des études supérieures ont des meilleures notes, et semblent plus motivés du fait que certaines variables tels que les sorties, internet… n'impactent pas négativement leurs résultats. Toutefois, ces résultats sont à prendre avec des pincettes au vu du nombre de répondants.

iii. Régression selon le fait de bénéficier de cours payant

De même que dans la partie précédente nous allons voir si ceux bénéficiant de cours payé le soir ont une meilleure motivation que ceux qui n'en n'ont pas. Pour ce faire, nous avons créé une nouvelle variable permettant de prendre en compte le fait qu'un élève ai ou non des cours du soir. La variable prend la valeur 1 si l'élève a au moins un cours et 0 sinon. Comme on peut le constater, cette nouvelle variable permet une meilleure séparation de notre échantillon car nous sommes

quasiment à 50 % dans les chaque groupes (paid-classe = 1 en orange et = 0 en bleu). Néanmoins, les individus n'ayant pas de cours du soir sont légèrement majoritaires.

a. Estimation de la moyenne de mathématiques

Intéressons-nous aux notes de mathématiques.

	TOUT L'ECHANTILLON	BENEFICIANT DE COURS DU SOIR (PAID-CLASS = 1)	NE BENEFICIANT DE COURS DU SOIR (PAID-CLASS = 0)
HIGHER	2.762	-1.474	1.279
PAID-CLASS	0.376	/	/
INTERNET	1.089	0.984	0.637
ROMANTIC	-0.715	-1.158	0.001
FREETIME	0.237	-0.091	0.076
GOOUT	-0.556	-0.237	-0.049
DALC	-0.181	-0.325	0.283
WALC	0.038	-0.007	-0.291

En moyenne, un étudiant bénéficiant de cours du soir obtient une note globale de 11.04 en maths tandis que les autres une note de 10.34. Une légère différence qui semble pencher du côté des élèves ayant du soutien, mais ils sont également légèrement plus nombreux dans l'échantillon. Etonnement, les variables qui peuvent impacter la motivation des étudiants comme être en couple, sortir, avoir du temps libre impactent négativement les étudiants ayant du soutien, ces variables impactent moins les étudiants n'ayant pas de cours du soir. Cela s'explique par le fait que les étudiants n'ayant pas de cours du soir peuvent ne pas en avoir besoin car il s'en sorte déjà bien cours (cela peut également être pour des raisons financières).

b. Estimation de la moyenne de portugais

	TOUT L'ECHANTILLON	BENEFICIANT DE COURS DU SOIR (PAID-CLASS = 1)	NE BENEFICIANT DE COURS DU SOIR (PAID-CLASS = 0)
HIGHER	1.24	0.597	1.279
PAID-CLASS	-0.531	/	/
INTERNET	0.28	-0.481	0.637
ROMANTIC	0.131	0.377	0.001

FREETIME	0.028	-0.089	0.076
GOOUT	-0.0485	0.023	-0.049
DALC	0.058	-0.196	0.283
WALC	0.112	0.114	-0.293

En moyenne, un étudiant bénéficiant de cours du soir obtient une note globale de 11.96 en portugais tandis que les autres une note de 12.37. Contrairement aux mathématiques, la meilleure moyenne penche du côté de ceux sans cours de soutien. On retrouve quasiment le même constat que pour les mathématiques, certaines variables jouent en défaveurs des ceux bénéficiant de soutien. Néanmoins, le fait d'être en couple et de sortir impacterait moins leurs résultats. On peut supposer que cela est dû au fait qu'il ne s'intéresse pas forcément à l'obtention des meilleurs résultats possibles.

En définitive, le fait d'avoir un soutien scolaire contrairement à ce que l'on pourrait penser n'influe pas autant que l'on pense sur des variables dites de loisirs. Au contraire, il semblerait que le fait de ne pas avoir de cours de soutien permettrait que les variables loisirs n'impactent pas négativement les notes.

Chapitre 2 : Discrimination sur le marché du travail

V. Introduction

Cette partie est dédiée à l'exploration de potentielles discriminations dans notre jeu de données. Pour rappel, une discrimination constitue un traitement défavorable injustifié envers certaines catégories d'une population, comme le sexe, l'ethnie, la religion ou la présence d'un handicap. Dans le cadre de notre thématique et de nos données, nous pouvons légitimement nous demander s'il existe des discriminations en fonction du genre, et le cas échéant quelles sont les caractéristiques qui y contribuent le plus.

Pour réaliser cette analyse, nous avons donc décidé d'étudier la présence de discriminations dans la notation finale de portugais, puis de mathématiques. Si tel est le cas, il peut être intéressant de constater si les mêmes variables impactent la discrimination pour l'une ou l'autre note, ou bien si cette dernière est plutôt influencée par des caractéristiques différentes en fonction de la matière à laquelle l'on s'intéresse.

Nous allons ainsi mobiliser la décomposition de Blinder-Oaxaca à l'aide du package *oaxaca* sur RStudio. Cette méthode, généralement utilisée pour étudier les discriminations sur le marché du travail (discrimination salariale notamment) selon le genre ou la nationalité, se prête tout aussi bien à notre thématique. En effet, cette méthode consiste à décomposer l'écart entre les résultats moyens de deux groupes (ces deux groupes étant déterminés à l'aide d'une variable à deux modalités mutuellement exclusives qui induit leur présence dans l'un ou l'autre groupe) et à répartir cet écart en une partie explicable par les différences de caractéristiques entre les deux groupes, et en une autre partie qui n'est pas imputable à ces différences de caractéristiques que l'on considérera donc comme inexpliquée et, par extension, engendré par de la discrimination.

Ainsi, au cours de ce chapitre, nous allons dans un premier temps réaliser quelques statistiques descriptives. Dans un second temps, nous utiliserons la méthode de la décomposition de Blinder-Oaxaca pour étudier les discriminations, puis nous conclurons.

VI. Stats Descriptives

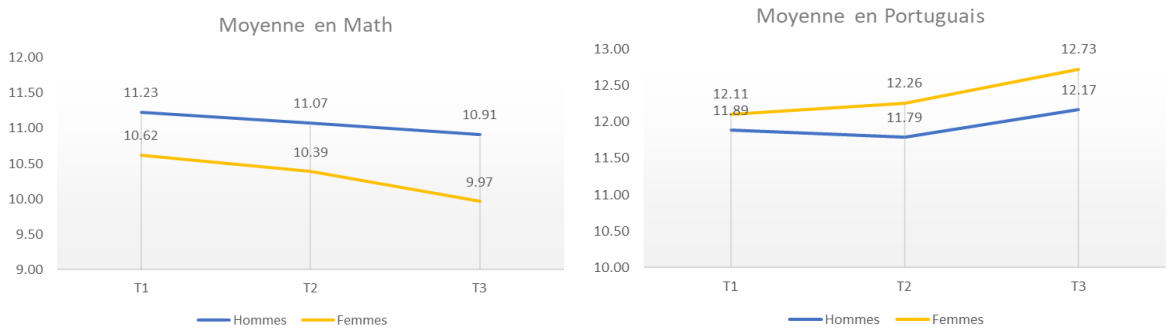
Pour commencer, nous avons réalisé quelques statistiques descriptives autour des différentes variables que nous avons utilisées pour mesurer une potentielle discrimination. Nous étudierons pour chaque variable, les moyennes scolaire, la consommation d'alcool, l'absentéisme, le nombre d'échecs, le nombre de sorties et le temps de travail personnel.

A. Inégalités hommes – femmes

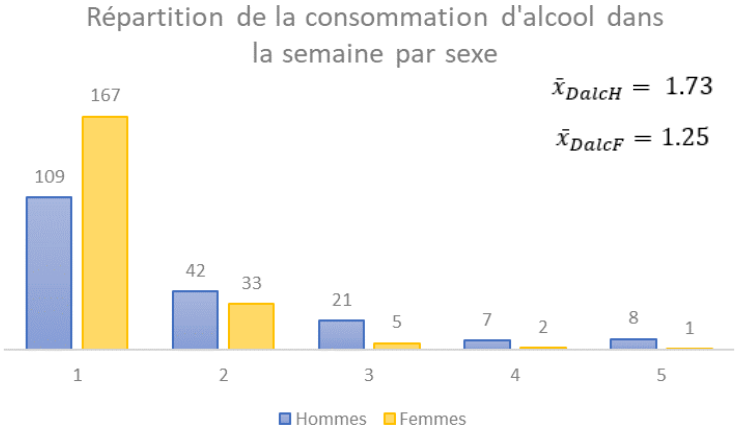
Nous avons décidé d'étudier la différence entre homme et femme car c'est l'une des inégalités les plus admise et documentée. Si l'on s'intéresse à un article du gouvernement sur l'égalité entre les

filles et les garçons, on apprend que les filles réussissent en moyenne plus que les garçons. Cependant elles s'orientent beaucoup moins en direction de filières scientifiques[2].

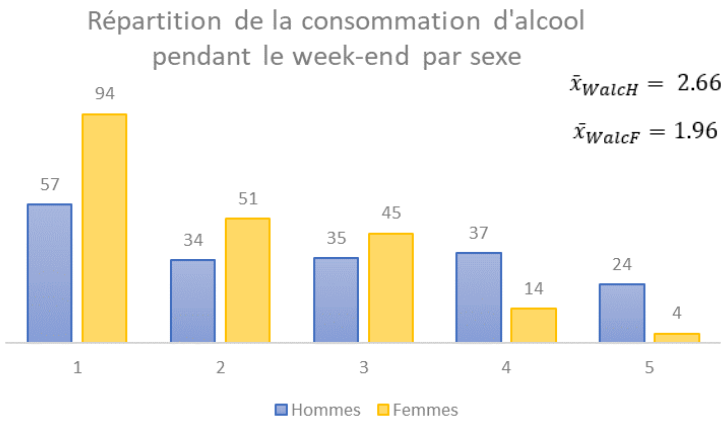
Pour commencer, notre échantillon est composé de 187 hommes et de 208 femmes, les garçons ont en moyenne une note en mathématiques plus élevée de 0.75 point par rapport aux filles. Cependant, on observe l'effet inverse lorsque l'on s'intéresse aux notes aux Portugais. En effet, les filles ont en moyenne de meilleure note que les garçons, cependant l'écart est plus faible (0.42 points).



Si l'on s'intéresse à la consommation d'alcool, les hommes sont les plus gros consommateurs, avec une plus grande part d'homme buvant beaucoup le week-end et un peu la semaine. De plus, près de la moitié des femmes (94) ne consomment pas ou de faibles quantités d'alcool pendant le weekend, alors c'est le cas pour seulement 1 tiers des garçons. En moyenne, le weekend, les garçons sont des consommateurs d'alcool de catégorie 3 (ce qui correspond à une consommation moyenne d'alcool) alors les filles sont des consommatrices de catégorie 2 (ce qui correspond à une faible quantité d'alcool)

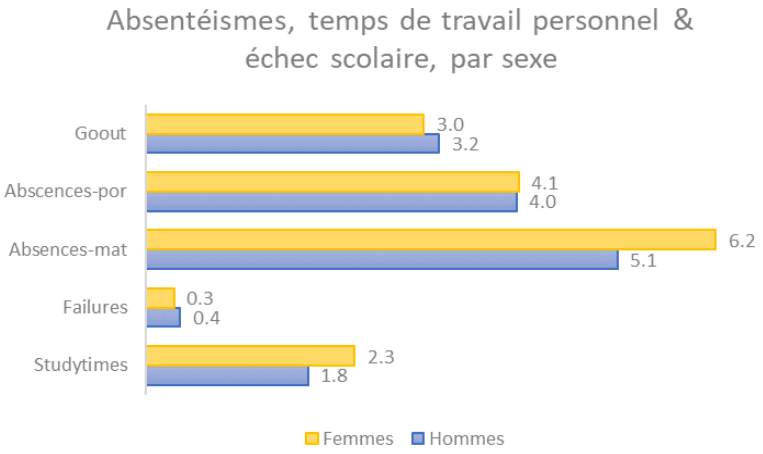


[2] « Égalité entre les filles et les garçons ». *Ministère de l'Education Nationale et de la Jeunesse*, https://www.education.gouv.fr/egalite-entre-les-filles-et-les-garcons-9047. Consulté le 11 juin 2023.



Nous pouvons également remarquer que les hommes ont tendance à avoir un comportement "problématiques" avec l'alcool en se rangeant dans la catégorie 4 et 5. Ils se définissent alors comme des gros consommateurs d'alcool. C'est le cas pour 61 d'entre eux contre 18 pour les femmes.

Sur le graphique ci-dessous, nous avons compilé de nombreuses informations, absence-por et absences-mat sont le nombre de jours d'absence, "goout" représente le nombre de sortie entre amis, failures le nombre de cours non validé.

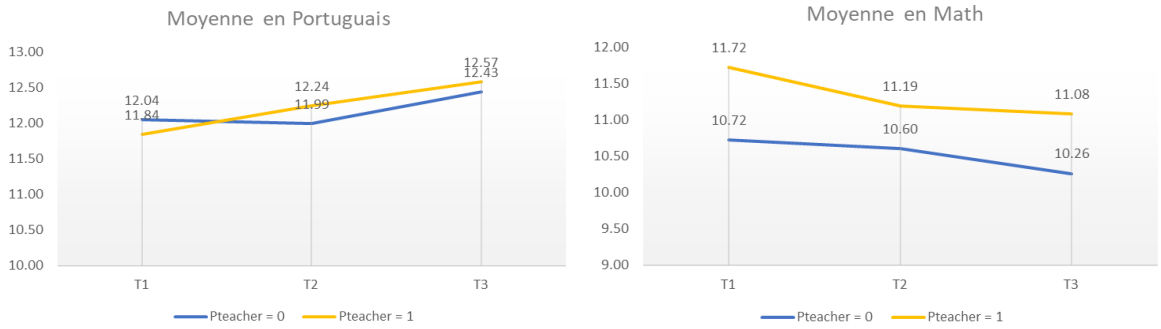


Nous pouvons ainsi constater que les hommes et les femmes ont relativement le même taux d'échec, d'absence en portugais et de sortie entre amis. Cependant, il y a un plus grand nombre d'absences en mathématiques en général. De plus, les femmes ont en moyenne 1,1 jours d'absences en plus par rapport aux garçons dans cette discipline. Pour finir, les femmes consacrent en moyenne plus de temps que les garçons à étudier, même si la différence reste faible.

B. Inégalités entre les enfants de professeur et les autres

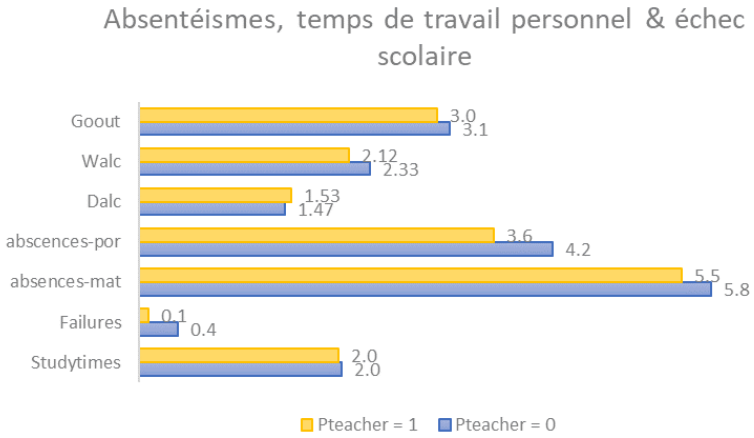
Pour cette seconde décomposition, nous avons choisi de découper l'échantillon en fonction de la variable "pteacher" qui est égale à 1 si l'étudiant à un parent professeur, 0 sinon. Il y a au total 19 % des étudiants qui ont au moins 1 de leurs parents professeurs, ce qui représente qu'une faible part de l'échantillon. Nous avons opéré les mêmes comparaisons que pour la précédente variable. Ainsi, dans un premier temps, nous allons comparer les moyennes des 2 groupes.

Sur les graphiques ci-dessous, nous pouvons observer les moyennes en mathématiques et en portugais :



Nous sommes en mesure de constater que les étudiants ayant un parent professeur ont de meilleures moyennes en math et des moyennes légèrement supérieures en portugais. Il y a en moyenne un écart de 0.8 point en math et un écart de 0.1 point en portugais entre les 2 groupes (toujours en faveur des étudiants ayant un parent professeur).

Le graphique ci-dessous reprend les mêmes éléments qu'auparavant en ajoutant la consommation d'alcool moyenne par semaine et par week-end.



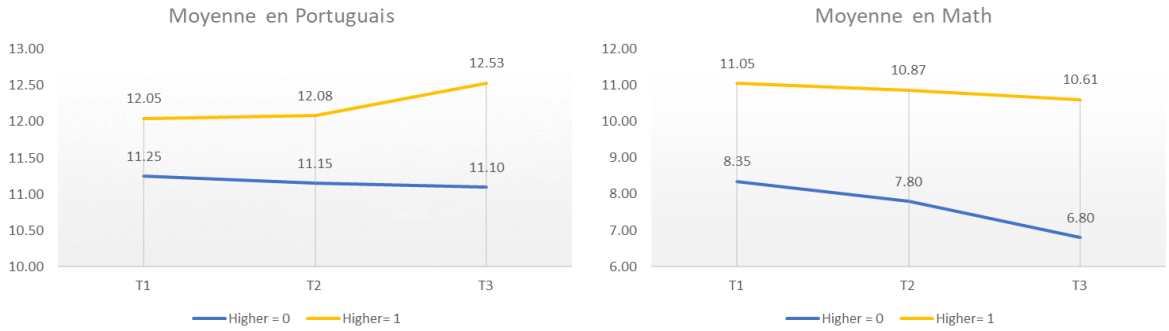
Dans un premier temps nous pouvons constater que le temps de travail personnel, la quantité de sortie sont similaire entre les 2 groupes. Concernant la consommation d'alcool, les étudiants ayant un professeur parmi leurs parents ont une moyenne de 1.53 par semaine (contre 1.47 pour les autres) et 2.12 en weekend (contre 2.33). Leur consommation est donc légèrement supérieure en semaine par rapport aux autres étudiants, mais reste inférieure durant le week-end. Il semble donc qu'avoir un de ses parents professeurs ne semble pas avoir d'effet significatif sur la consommation d'alcool des étudiants. Ensuite, nous remarquons que les enfants de professeurs ont une probabilité plus faible d'être absent, en math ou en portugais, même si l'écart est plus faible.

La différence la plus flagrante concerne la variable failures, où les parents de professeurs n'ont qu'une très faible probabilité de ne pas valider une matière par rapport à la moyenne.

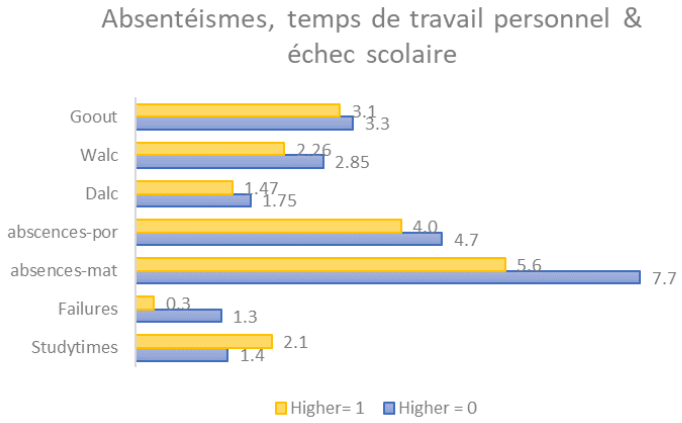
C. Inégalité entre ceux voulant suivre des études et les autres

Cette troisième décomposition ne peut pas être définis comme une discrimination, car la volonté de poursuivre ou non ses études est un choix personnel de l'étudiant. Cependant, dans une seconde partie nous allons chercher à estimer, via un modèle Probit, cette variable en fonction de différente variable comme le sexe, l'environnement familiale etc… afin de voir si ce choix personnel est lié ou non à des caractéristiques socio-économiques.

Il n'y a que 20 étudiants ne voulant pas poursuivre d'études supérieures (contre 375) ce qui représente une part très faible de notre échantillon. Ainsi, c'est dans cette décomposition que les écarts de moyenne observées sont les plus élevés. En effet, il y a un écart moyen de 3.19 points en mathématique et un écart moyen de 1.05 point en portugais.



Les étudiants ne voulant pas suivre leurs études sont également ceux qui connaissent une plus grande baisse de moyenne au cours de l'année scolaire. Sur le graphique ci-dessous, l'ensemble des variables tendent vers une même conclusion.



Les étudiants ne voulant pas suivre d'études supérieures sont ceux qui étudient le moins, sont le plus absents, boivent le plus d'alcool toutes périodes confondues et qui sortent le plus. De plus, ce sont eux qui échouent le plus avec 1.3 matières non validées en moyenne.

D. Autres inégalités

Dans cette dernière partie, nous avons ajouté des variables pour lesquelles nous avons fait des tests, les résultats seront disponibles en annexes.

Dans un premier temps, nous avons utilisé la variable “rural”. Pour justifier ce choix, nous disposons de l’ouvrage de Christophe Guilluy : “La France Périphériques”[3], qui est un livre revenant sur les fragilités sociales et économiques des personnes vivant dans les régions rurales.

Les résultats pour cette variable ne sont pas concluants. En effet, même si les étudiants “ruraux” ont une moyenne légèrement inférieure en mathématiques (0.95 points), ils ont des notes presque identiques en portugais. Concernant les autres variables, les ruraux sont en moyenne de plus gros consommateurs d’alcool, plus absents en cours. Cependant, il n’y a pas d’écart notable pour le reste des variables.

L’ensemble de ces éléments nous ont amenés à ne pas réutiliser cette variable pour la partie suivante.

Dans un second temps, nous avons utilisé la variable “GP” qui correspond à l’appartenance à une certaine école. L’objectif de cette comparaison était d’étudier s’il y avait une favorisation des étudiants d’une école à une autre. Cependant, à cause du faible nombre de personne étant dans l’école Mousinho da Silveira (50), et de résultats faibles, nous doutons de la pertinence de cette variable. La seule chose notable est que la moyenne en math est supérieure dans l’école Gabriel Perreira alors qu’il y a beaucoup plus d’absence en math dans cette même école.

VII. Méthodes de la décomposition de Blinder-Oaxaca

A. Discriminations sexuelles

Dans un premier temps, nous avons décidé d’explorer les potentielles discrimination en fonction du genre. La variable *woman*, qui pour rappel prend la valeur 0 quand l’individu est un homme et la valeur 1 s’il s’agit d’une femme, nous servira donc à scinder notre jeu de données pour former les deux groupes. A cet effet, nous pouvons revenir sur la composition de nos données pour observer la taille de nos deux groupes.

```
> results$n
$n.A
[1] 187

$n.B
[1] 208

$n.pooled
[1] 395
```

L’extrait de code ci-dessus, extrait de la sortie de RStudio, nous indique que parmi les 395 individus composant notre jeu de données, respectivement 187 hommes et 208 femmes ont été recensés. La somme de ces deux valeurs étant égale à la taille de notre échantillon initial, nous pouvons en conclure que le sexe de chaque individu du jeu de données a été renseigné. Si cela

[3] Guilluy, Christophe. « Introduction ». *La France périphérique*, Flammarion, 2015, p. 7-12. *Cairn.info*, https://www.cairn.info/la-france-peripherique--9782081347519-p-7.htm.

n'avait pas été le cas, les individus dont la caractéristique déterminante de l'appartenance à l'un ou l'autre groupe est absente n'auraient tout simplement pas été pris en compte dans l'étude.

La suite de notre analyse, réalisée de façon indépendante sur les notes de portugais puis sur les notes de mathématiques, sera donc réalisée sur l'intégralité de notre échantillon.

i. Moyenne en portugais

Dans un premier temps, pour décomposer l'écart de notes finales de portugais entre hommes et femmes, nous devons construire un modèle contenant plusieurs spécifications, présenté ci-dessous.

```
Code : oaxaca (formula = G3por ~ failures + Medu + goout + Pteacher + GT3 + schoolsup + studytime + rural + famsup + higher + absencespor + paidpor | woman | Pteacher, data = data, R=500)
```

Cette équation présente plusieurs composantes. Dans un premier temps, nous spécifions la variable à expliquer, dans notre cas la note finale de portugais. Nous ajoutons ensuite au modèle les variables exogènes dont les caractéristiques pourraient avoir un effet causal sur notre variable expliquée. Nous isolons ensuite la variable qui détermine l'appartenance aux deux groupes. Enfin, nous indiquons également la ou les variables binaires que l'on souhaite inclure dans l'analyse, pour examiner comment elles contribuent aux différences de notes. Dans notre cas, nous avons ajouté la variable Pteacher, qui prend la valeur 1 si l'élève a au moins un parent enseignant, et 0 sinon. Nous décidons également de procéder à du bootstrap, autrement dit répliquer l'échantillon un nombre de fois défini, ici 500.

Après avoir présenté l'équation principale de notre modèle, nous pouvons à présent nous intéresser aux écarts de notes entre hommes et femmes.

```
> results2$y
$y.A
[1] 12.16578

$y.B
[1] 12.72596

$y.diff
[1] -0.5601861
```

La sortie de code ci-dessus nous permet de constater que la note moyenne masculine pour l'examen final de portugais est d'environ 12,17/20 contre 12,73/20 pour les femmes. La différence entre les deux groupes s'élève à environ -0.56.

A première vue, cela pourrait nous étonner car historiquement, les discriminations avérées sont davantage dirigées envers les femmes plutôt que les hommes. Mais d'un autre côté, il est souvent admis que les filles ont de meilleures capacités que les garçons particulièrement dans les matières littéraires. Contre toute attente, nous nous dirigeons dans un premier temps vers l'étude d'une potentielle discrimination envers les hommes.

```
> results2$threefold$overall
  coef(endowments)     se(endowments) coef(coefficients)   se(coefficients)
         0.2589417          0.1937179         -0.5928732          0.2684594
 coef(interaction)    se(interaction)
        -0.2262546          0.2664860
```

La sortie de code ci-dessus est le résultat d'une décomposition en trois parties. Nous distinguons ainsi la partie *endowment*, la partie *coefficient* et la partie *interaction*, auxquelles étant attribuée une partie de la différence moyenne entre nos deux groupes relevés précédemment. Autrement dit, la somme des trois coefficients correspond à l'écart de note moyen entre hommes et femmes.

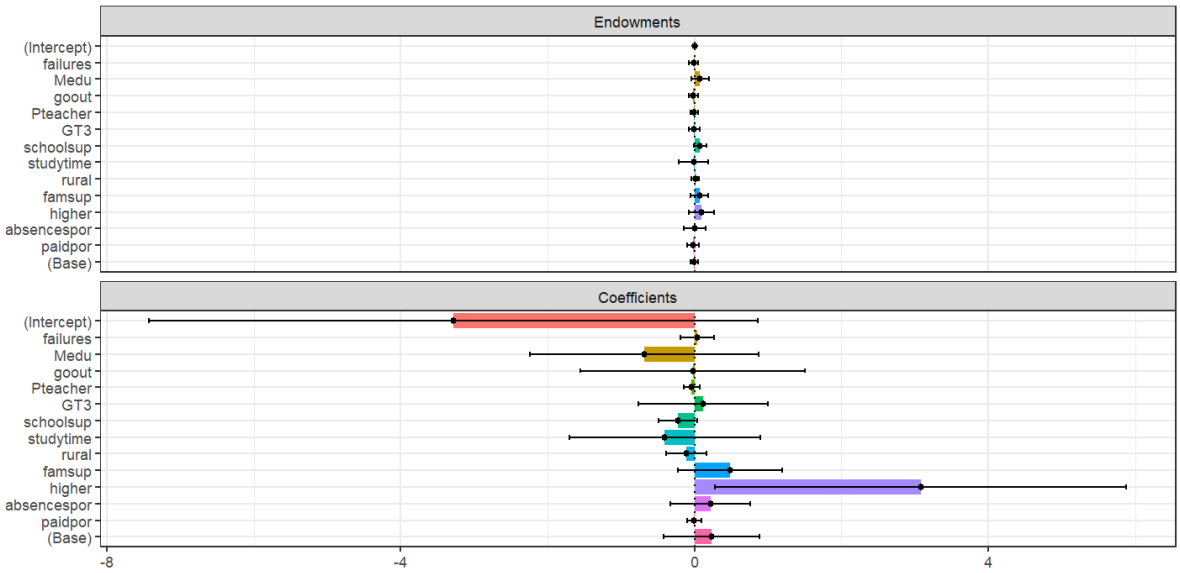
Le coefficient associé à l'*endowment* correspond à la part de l'écart imputable à des différences de caractéristiques entre les deux groupes, il s'agit donc de la partie de l'écart que nous sommes en mesure d'expliquer. Le coefficient endowment s'élève à environ 0.26, ce qui est surprenant dans la mesure où nous décomposant une valeur négative. En d'autres termes, les différences de caractéristiques entre les hommes et les femmes tendent à resserrer cet écart plutôt qu'à le creuser.

Le coefficient associé à la partie *coefficient* correspond au rendement de ces caractéristiques, il s'agit donc de la partie inexplicable de l'écart de note moyen qui s'élève à -0.59. Cela implique que les rendements de caractéristiques contribuent positivement à l'écart de notes observé en exacerbant la différence de moyennes entre les groupes.

Enfin, le coefficient d'interaction est de l'ordre de -0.23 environ, ce qui signifie qu'il contribue également à creuser l'écart de notes entre hommes et femmes.

Le graphique ci-dessus reprend les variables de notre model initial en considérant les parties *endowment* et *coefficient*s explicités précédemment. Dans le premier graphique, l'on devrait être en mesure de déterminer la variable qui explique la plus grande partie de l'écart induit par les différences de caractéristiques entre les deux groupes, mais aucune variable ne semble

particulièrement se distinguer vis-à-vis des autres.



Dans le graphique associé aux coefficients, la variable expliquant la plus grande partie de l'écart inhérente aux différences de caractéristiques est sans conteste la variable *higher* qui prend la valeur 1 si l'étudiant à l'intention de poursuivre ses études dans le supérieur, 0 sinon. Ainsi, ce sont les rendements de la motivation à poursuivre des études qui impactent le plus l'écart de notes observé entre hommes et femmes.

```
> results2$beta$beta.diff["higher"]
  higher
3.141008
```

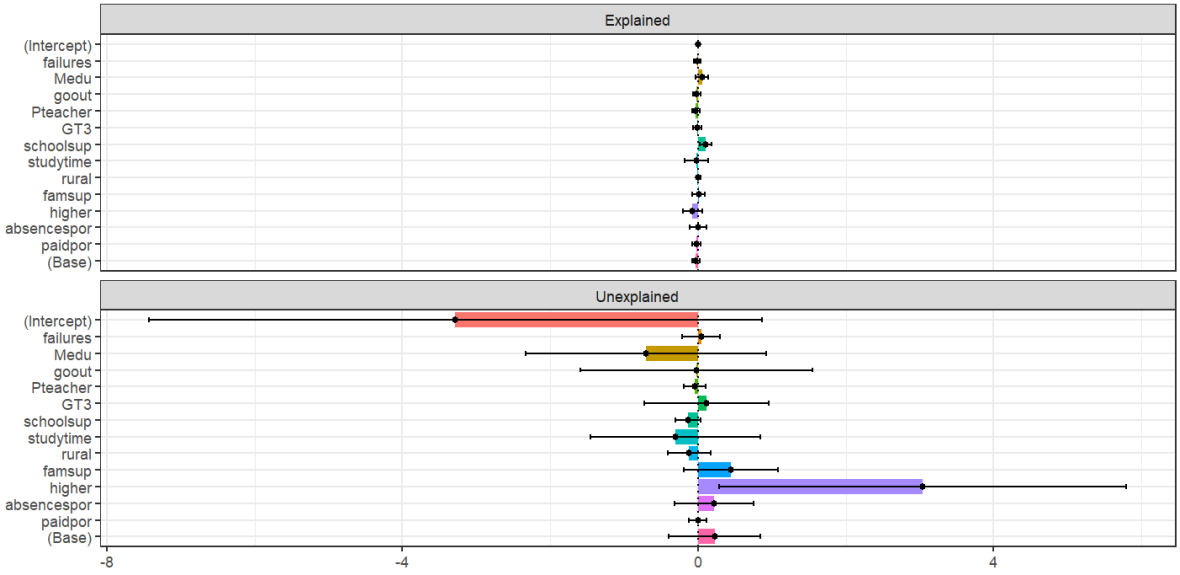
Le résultat ci-dessus nous indique que le rendement marginal de l'intention de poursuite d'études est plus élevé chez les hommes de 3.14 points par rapport aux femmes.

Pour la suite de notre étude, nous allons cette fois-ci procéder à une décomposition en deux parties, en négligeant le terme d'interaction pour ne conserver que l'*endowment* et le *coefficient*. Pour ce faire, nous allons repartir de l'écart moyen initialement observé entre nos deux groupes, et le décomposer en une partie expliquée et une partie non-expliquée. C'est l'objet de la sortie de code ci-dessous.

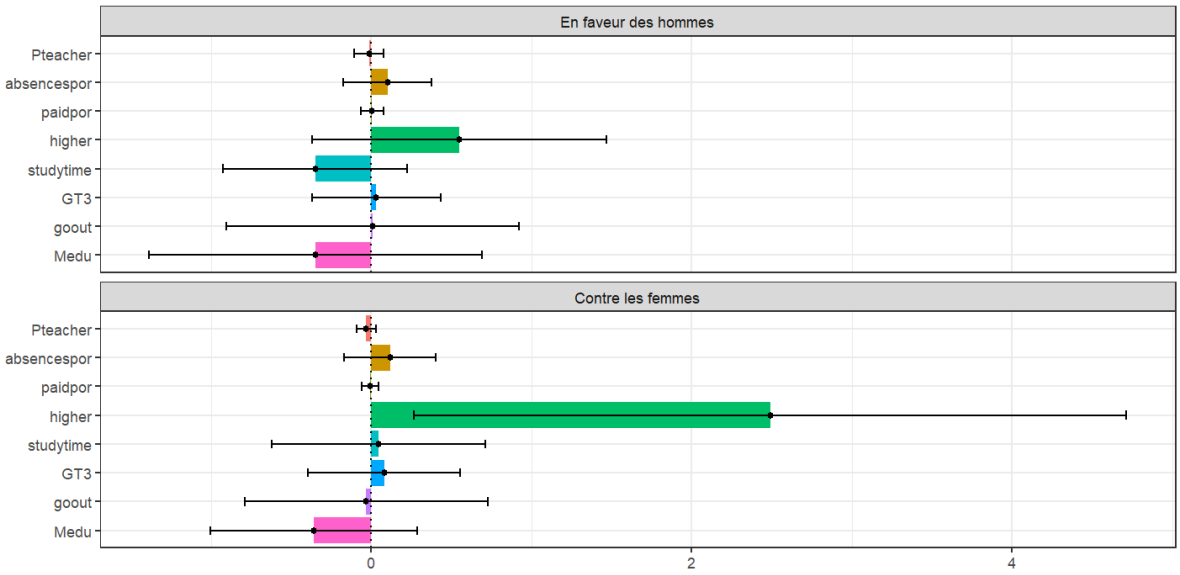
```
> results2$twofold$overall
     group.weight coef(explained) se(explained) coef(unexplained) se(unexplained)
[1,]    0.0000000      0.25894170     0.1937179        -0.8191278       0.3037389
[2,]    1.0000000      0.03268705     0.2004080        -0.5928732       0.2684594
[3,]    0.5000000      0.14581438     0.1452285        -0.7060005       0.2537917
[4,]    0.4734177      0.15182874     0.1456420        -0.7120149       0.2528315
[5,]   -1.0000000     -0.03818296     0.1473276        -0.5220032       0.2107696
[6,]   -2.0000000      0.06375646     0.1478365        -0.6239426       0.2568136
     coef(unexplained A) se(unexplained A) coef(unexplained B) se(unexplained B)
[1,]       -8.191278e-01      3.037389e-01           0.0000000         0.0000000
[2,]        0.000000e+00      0.000000e+00          -0.5928732         0.2684594
[3,]       -4.095639e-01      1.518695e-01          -0.2964366         0.1342297
[4,]       -4.313382e-01      1.437954e-01          -0.2806767         0.1413660
[5,]       -2.748776e-01      1.118689e-01          -0.2471256         0.1005653
[6,]       -5.925318e-14      1.545079e-14          -0.6239426         0.2568136
```

Nous considérons ici la cinquième ligne du tableau. Il s'agit de considérer l'ensemble de notre échantillon en ne conservant pas la variable *woman* qui nous a permis de constituer nos groupes. Les autres lignes sont sujettes à d'autres pondérations, mais les résultats observés convergent avec ceux de la cinquième ligne que nous allons présenter.

L'écart de notes moyennes de -0.56 est donc décomposé en une partie expliquée (-0.03) et une partie non-expliquée (-0.52). Nous pouvons ainsi constater qu'avec cette méthode de décomposition, la grande majorité de l'écart n'est pas expliquée par des différences de caractéristiques entre nos deux groupes, mais bien par le rendement de ces caractéristiques. De plus, la partie inexpliquée est elle-même décomposée en deux coefficients : une partie inexpliquée en faveur des hommes (-0.27) et une partie inexpliquée en défaveur des femmes (-0.25).

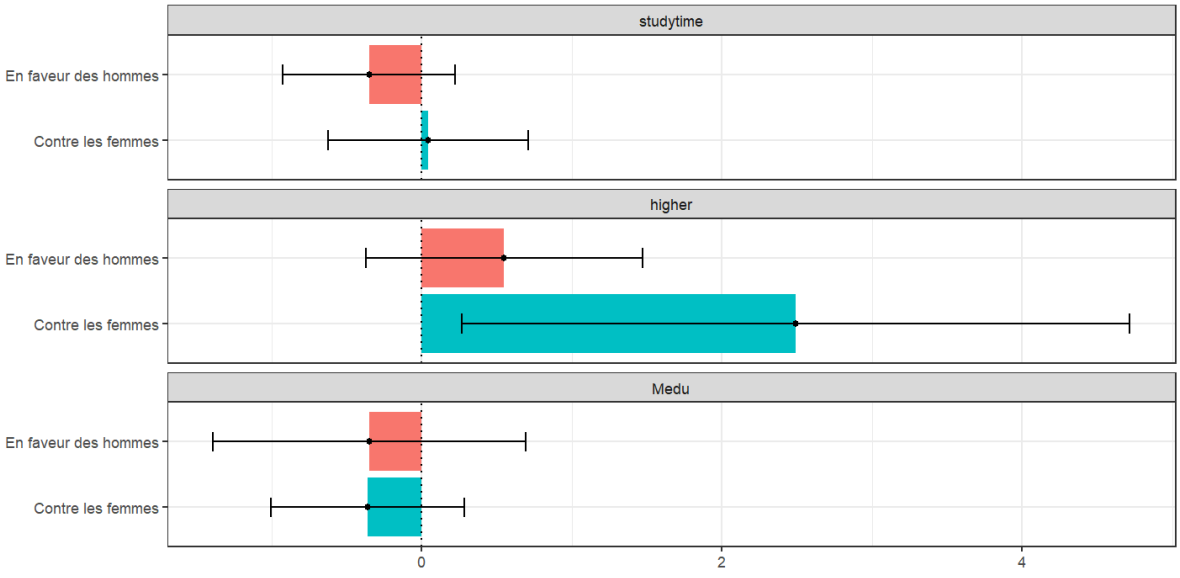


Le graphique ci-dessus présente des résultats similaires à ceux obtenus dans le cadre de la décomposition en trois parties : si aucune variable ne semble se démarquer dans l'explication de la part expliquée de l'écart de notes entre nos deux groupes, la variable *higher* contribue hautement à la part inexpliquée de cet écart. Comme la partie inexpliquée de l'écart de notes moyen a elle-même été décomposé entre deux parties, une en faveur des hommes et une en défaveur des femmes, nous pouvons également mettre en lumière les variables qui contribuent à creuser l'écart entre les deux groupes dans un sens ou dans l'autre.



Comme nous décomposons un écart négatif, la lecture du graphique ci-dessus est légèrement altérée. En effet, le fait de vouloir faire des études supérieures (*higher*) joue plus en faveur des femmes qu'il ne joue en défaveur des hommes. Le niveau d'éducation de la mère (*Medu*) joue tout autant en faveur des hommes qu'il ne joue en défaveur des femmes. Enfin, le temps de travail personnes (*studytime*) joue davantage en faveur des hommes qu'il ne joue en faveur des femmes.

Ces trois relations sont observables de façon plus claire dans les graphiques ci-dessous, et il est possible de leur attribuer des coefficients précis avec le tableau associé.



```
> results2$twofold$variables[[5]][variables2, columns2]
          group.weight coef(unexplained A) coef(unexplained B)
studytime           -1          -0.3494528          0.04588526
higher              -1           0.5498760          2.49084173
Medu                -1          -0.3475781         -0.35802006
```

Pour conclure cette première analyse qui concernait l'écart de note moyen entre hommes et femmes pour l'examen final de portugais, force est de constater que nos résultats sont assez préoccupants. En effet, nous avons dans un premier temps mis en lumière un écart de moyennes négatif entre femmes et hommes au détriment de ces derniers. En décomposant cet écart en trois parties, nous avons pu constater que les différences de caractéristiques imputables à nos deux groupes tendaient à resserrer l'écart observé plutôt qu'à le creuser, mais qu'aucune variable ne se démarquait pour l'expliquer. D'un autre côté, l'étude de la partie inexpliquée de l'écart a pu mettre en valeur l'importance de la motivation à poursuivre des études dans l'explication de ce phénomène. Une analyse plus poussée nous a permis de déterminer que cela joue davantage en faveur des femmes qu'en défaveur des hommes. D'autres caractéristiques telles que le niveau d'éducation de la mère et le temps de travail personnel rentrent également en compte pour appréhender cet écart inexpliqué. Néanmoins, au vu des interprétations qui en découlent, chaque variable a un effet très singulier, ce qui rend nos conclusions très confuses. Nous pouvons néanmoins retenir qu'une part de la différence de notes moyennes entre femmes et hommes reste inexpliquée.

Notre objectif initial était de déterminer s'il existe des différences injustifiées de notation entre les femmes et les hommes, nous pouvons à présent voir si les analyses que nous avons réalisées précédemment coïncident avec l'analyse d'une autre note, si une part inexpliquée demeure toujours, et le cas échéant si d'autres variables rentrent en jeu pour expliquer cette part inexpliquée de la différence de moyennes.

ii. Moyenne en mathématique

Les individus composant notre jeu de données et la caractéristique mise en avant pour la construction des deux groupes restant inchangés pour l'analyse des écarts de moyenne de mathématiques, la taille et la composition des groupes homme et femme sont donc identiques à celles des groupes utilisés pour analyser la différence de moyennes pour la note finale de portugais. Nous pouvons tout de même rappeler la répartition dans les deux groupes telles qu'elle avait été réalisée à l'origine.

```
> results$n
$n.A
[1] 187

$n.B
[1] 208

$n.pooled
[1] 395
```

```
Code : oaxaca (formula = G3mat ~ failures + Medu + goout + Pteacher +
GT3 + schoolsup + studytime + rural + famsup + higher + absencesmat +
paidmat | woman | Pteacher, data = data, R=500)
```

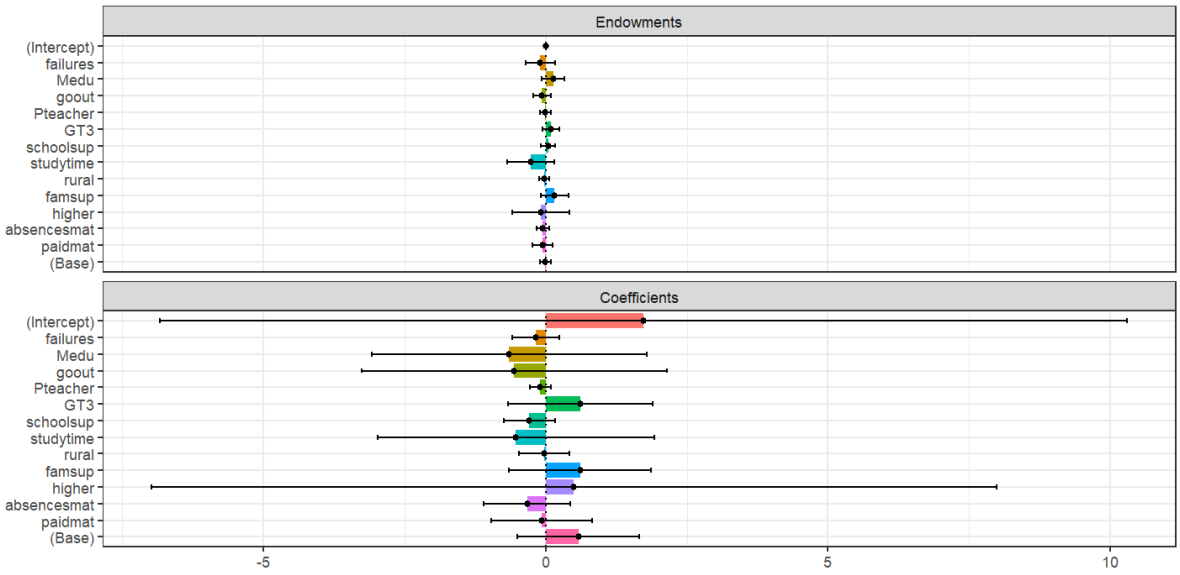
Nous retrouvons également l'équation nous permettant de construire le modèle destiné à décomposer l'écart de moyenne entre les deux groupes. La variable expliquée correspond à présent à la note finale d'examen de mathématiques, et les variables liées aux absences et aux cours particuliers sont maintenant associées à l'enseignement des mathématiques. Nous pouvons à présent évaluer la différence de note moyenne existante entre nos deux groupes, puis la décomposer pour étudier la possible présence d'une part non-expliquée.

```
> results$y
$y.A
[1] 10.91444

$y.B
[1] 9.966346

$y.diff
[1] 0.9480923
```

Cette fois-ci, nous sommes en mesure de constater que la moyenne pour l'examen final de mathématiques est d'environ 10.91/20 pour les hommes et 9.95/20 pour les femmes. Ainsi, nous nous retrouvons dans ce cas avec une moyenne plus faible chez les femmes de l'ordre de 0.95 points par rapport aux hommes. Nous allons une fois de plus décomposer cet écart en trois parties *endowments*, *coefficients* et *interaction* et voir quelle part de l'écart est imputée aux différentes parties.

```
> results$threefold$overall
  coef(endowments)     se(endowments) coef(coefficients)   se(coefficients)
        -0.2327502          0.4814098          1.3222262          0.4839874
 coef(interaction)    se(interaction)
        -0.1413837          0.5570944
```

La décomposition en trois parties de la différence des moyennes de notes nous permet dans un premier temps d'observer que les différences de caractéristiques inhérentes aux deux groupes tendent à resserrer l'écart entre leurs moyennes. D'autre part, les rendements de ces caractéristiques soit la part inexpliquée de l'écart contribuent pour leur part à davantage creuser l'écart de notes. Enfin, le coefficient d'interaction tend au même titre que l'*endowment* à résorber l'écart de notes moyennes existant entre les deux groupes.



Le graphique ci-dessus illustre les résultats de la première décomposition en omettant le coefficient d'interaction. De même que pour les moyennes de portugais, aucune variable ne semble spécialement se distinguer dans l'explication de la part de la différence de moyennes imputables aux différences de caractéristiques entre les hommes et les femmes. Nous avons néanmoins choisi de nous intéresser au temps de travail personnel *studytime* dans la mesure où il se démarque légèrement des autres variables. Cela nous permettra dans le même temps d'interroger sa significativité pour voir si son impact sur cette partie de l'écart de moyennes est avéré.

```
> summary(results$reg$reg.pooled.2)$coefficients["studytime",]
  Estimate Std. Error    t value   Pr(>|t|)
 0.4157312  0.2693155  1.5436586  0.1235012
```

Dans un premier temps, les résultats ci-dessus que les femmes révisent davantage que les hommes, toutes chose étant égales par ailleurs, néanmoins l'analyse de la p-value nous indique que ce coefficient n'est pas significatif même au seuil de 10%. Nous pouvons donc en conclure que le temps de travail personnel a un effet non-significatif lorsqu'il s'agit d'expliquer la part de la différence de moyennes correspondant aux différences de caractéristiques entre les deux groupes.

```
> results$x$x.mean.diff["studytime"]
 studytime
-0.5141403
```

Nous pouvons néanmoins tirer d'autres informations de cette variable, même si son coefficient s'avère non-significatif. En effet, la sortie de code ci-dessus et plus spécifiquement le coefficient négatif associé à la variable nous permet de conclure que lorsque le temps de travail personnel est plus faible, les notes (ici spécifiquement la note finale de mathématiques) le sont également.

```
> results$beta$beta.diff["GT3"]
      GT3
0.8241159
```

Nous pouvons également nous intéresser à la variable GT3 qui rend compte de la composition de la famille en prenant la valeur 1 si l'individu a au moins deux frères ou sœurs. La sortie de code ci-dessus nous permet de constater que le rendement marginal de cette situation est plus élevé chez les femmes de 0.82 points par rapport aux hommes.

Enfin, si nous revenons sur le graphique ci-dessus pour analyser cette fois-ci la part de la différence de moyennes associé aux *coefficients*, autrement dit le rendement des caractéristiques inhérentes à chaque groupe qui correspond ainsi à la part non-expliquée de cette différence, nous pouvons constater que contrairement à ce qui avait été déterminé dans l'analyse des notes de portugais, aucune variable ne semble non-plus déterminante pour expliquer la part non-expliquée de l'écart de moyennes, néanmoins leur effet semble déjà plus impactant que dans le cadre des *endowments*.

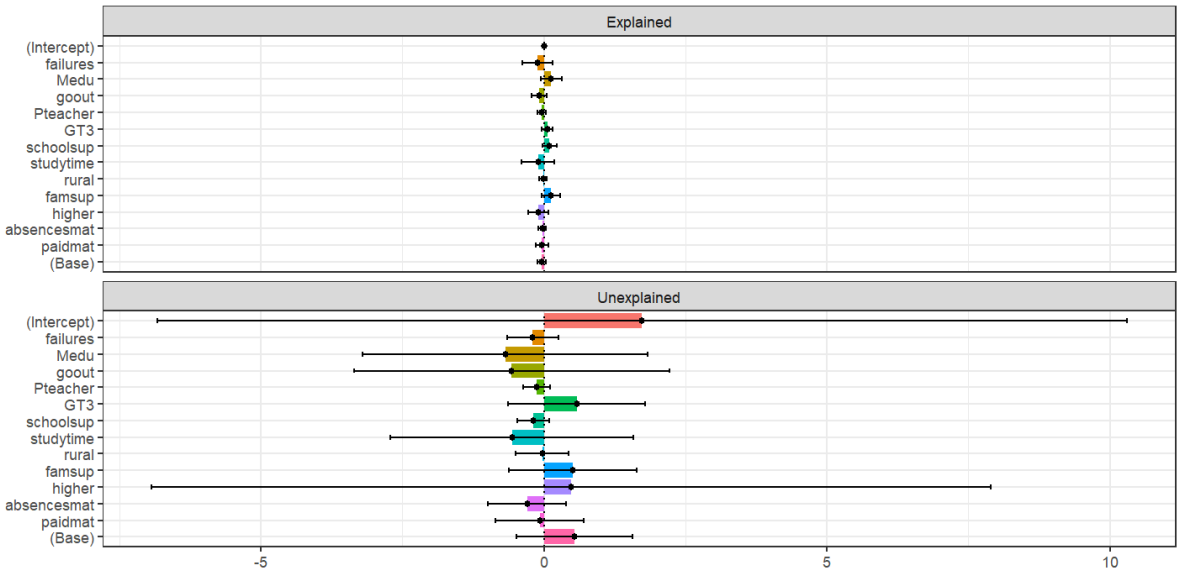
Nous allons tâcher de procéder à une décomposition en deux parties seulement, en ne conservant que l'endowment et le *coefficient*, et ainsi étudier la part expliquée et non-expliquée de l'écart des moyennes de mathématiques entre hommes et femmes.

```
> results$twofold$overall
     group.weight coef(explained) se(explained) coef(unexplained) se(unexplained)
[1,]    0.0000000      -0.2327502     0.4814098          1.180843       0.6220784
[2,]    1.0000000      -0.3741339     0.3986376          1.322226       0.4839874
[3,]    0.5000000      -0.3034421     0.3431401          1.251534       0.4827257
[4,]    0.4734177      -0.2996838     0.3406289          1.247776       0.4787309
[5,]   -1.0000000      -0.1554613     0.2799894          1.103554       0.3715858
[6,]   -2.0000000      -0.3851324     0.2949987          1.333225       0.4600777
     coef(unexplained A) se(unexplained A) coef(unexplained B) se(unexplained B)
[1,]        1.180843e+00      6.220784e-01           0.0000000         0.0000000
[2,]        0.000000e+00      0.000000e+00           1.3222262         0.4839874
[3,]        5.904213e-01      3.110392e-01           0.6611131         0.2419937
[4,]        6.218108e-01      2.945029e-01           0.6259653         0.2548592
[5,]        5.811118e-01      1.956129e-01           0.5224419         0.1794925
[6,]       -4.658773e-14      1.253714e-14           1.3332247         0.4600777
```

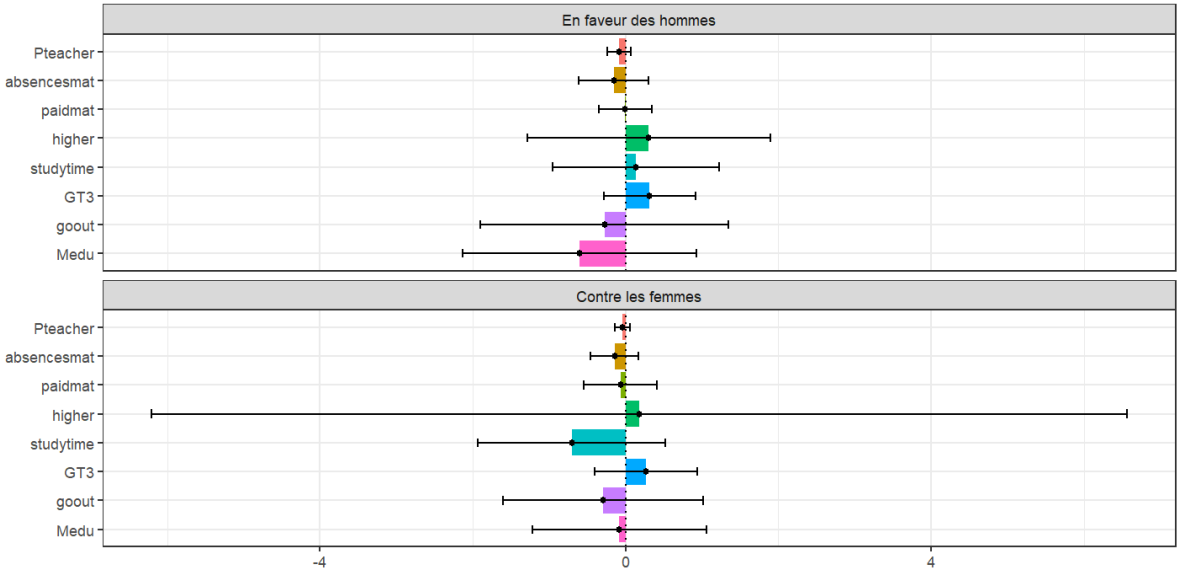
De même que pour l'analyse précédente, nous focalisons notre attention sur la cinquième ligne qui correspond à l'ensemble de notre échantillon en omettant la variable *woman* ayant servi à déterminer l'appartenance aux deux groupes. Les résultats sont similaires avec ceux obtenus dans le cadre de la première décomposition dans la mesure où les différences de caractéristiques entre les deux groupes tendent à résorber l'écart entre les moyennes, tandis que les rendements de ces caractéristiques (soit la partie restant inexpliquée) contribuent à creuser l'écart entre nos deux groupes. Il semblerait ainsi que l'intégralité de l'écart de notes demeure inexpliquée, ce qui semble relativement incohérent mais nous prendrons le temps de discuter l'intégrité de nos résultats en conclusion.

Si l'on décompose à présent cette partie inexpliquée, l'écart de 1.10 se découpe en deux parties : un coefficient de 0.58 en faveur des hommes et de 0.52 au détriment des femmes.

Nous pouvons poursuivre en analysant graphiquement le fruit de ces deux décompositions successives, par le biais des divers graphiques ci-après.

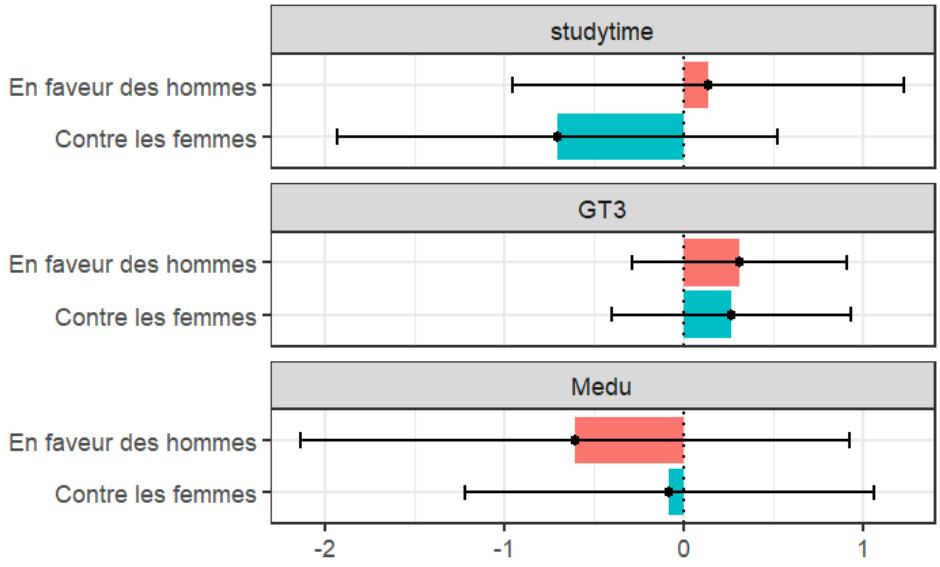


Dans un premier temps, nous retrouvons un graphique assez similaire à celui généré dans le cadre de la décomposition en trois parties. En effet, si aucune variable ne se démarque dans l'interprétation de l'écart expliqué par des différences de caractéristiques entre les deux groupes, nous pouvons faire le même constat pour les rendements de ces caractéristiques, même si leur effet est un peu plus important dans cette situation. Nous décidons tout de même de nous focaliser sur certaines variables dans le cadre de la décomposition de la partie inexpliquée.



Le graphique ci-dessus illustre les résultats de la décomposition de la partie inexpliquée de l'écart de moyennes entre les deux groupes. Nous pouvons notamment constater que le temps personnel de révisions (*studytime*) joue plus en défaveur des femmes qu'il ne joue en faveur des hommes. Le fait d'être dans une famille de trois enfants ou plus (*GT3*) joue presque autant en faveur des femmes qu'il ne joue en faveur des hommes. Enfin, le niveau d'études de la mère (*Medu*) joue davantage en défaveur des hommes qu'il ne joue en faveur des femmes.

Pour plus de précision, nous pouvons retrouver les visualisations graphiques des trois interprétations précédentes ainsi que les coefficients associés à chaque mesure dans le tableau associé ci-dessous.




```
> results$twofold$variables[[5]][variables, columns]
          group.weight coef(unexplained A) coef(unexplained B)
studytime           -1           0.1390253         -0.70578434
GT3                 -1           0.3120851          0.26792743
Medu                -1          -0.6047064         -0.07996608
```

Arrivés au terme de notre seconde analyse, nous pouvons à présent conclure quant aux résultats précédemment obtenus puis procéder à une comparaison entre les décompositions de l'écart moyen des deux matières. Dans un premier temps, nous avons mis à jour un écart de notes en défaveur du groupe composé de femmes, et nous avons pu constater en le décomposant en trois parties que les différences de caractéristiques entre les groupes contribuaient à resserrer l'écart de notes plutôt qu'à le creuser, et la même conclusion s'impose à nous dans le cadre de la décomposition en deux parties. Nous avons également pu mettre en évidence qu'aucune de nos variables n'avait de fort impact pour expliquer la part inexpliquée de l'écart de moyenne constaté entre nos deux groupes, mais que plusieurs variables comme le temps de révisions, la composition de la famille (nombre de frères et sœur) ou encore le niveau d'éducation de la mère y contribuaient de façon modérée.

VIII. Conclusion

Pour conclure, nous pouvons revenir sur les convergences ou les différences en termes de résultats quant aux deux matières différentes étudiées. Tout d'abord, nous avons observé un écart en faveur des femmes pour la note de finale de portugais et un écart en faveur des hommes pour la note finale de mathématiques. L'objet de notre analyse s'est donc constitué autour de l'étude d'une potentielle discrimination dans un sens puis dans l'autre. Dans les deux cas, nous avons pu constater que les différences de caractéristiques entre les deux groupes (*endowment*) tendaient à resserrer l'écart de notes plutôt qu'à l'aggraver, et qu'une grande part voire la totalité de l'écart demeurait inexpliquée dans certaines décompositions. Ce constat interroge dans la mesure où les notes dépendent nécessairement de l'implication en classe (temps de révisions, absences, cours particuliers), ce qui nous pousse à remettre en question la qualité de nos données. De plus, si nous avions pu constater la pertinence de la motivation à poursuivre des études dans l'explication de la part inexpliquée de l'écart de moyennes de portugais, aucune variable ne se démarque particulièrement pour interpréter la part inexpliquée de l'écart de notes en mathématiques. Cela remet une fois de plus en cause l'intégrité de nos données dans la mesure où, bien que l'on ait pu mettre en évidence que la majorité voire totalité de l'écart entre les groupes était inexpliqué, nous n'avons pas été en mesure de mettre en lumière des variables qui contribuent à expliquer cet écart. Une des ouvertures potentielles serait de réaliser un modèle Probit expliquant la variable « higher » afin d'expliquer ce qui détermine le choix des étudiants de poursuivre des études supérieures.

IX. Bibliographie

[1] Chiappori, Pierre-André. « Modèles d'appariement en économie. Quelques avancées récentes ». *Revue économique*, vol. 63, n° 3, 2012, p. 437-52. *Cairn.info*, https://doi.org/10.3917/reco.633.0437.

[2] Cortez, Paulo, et Alice Silva. « Using data mining to predict secondary school student performance ». *EUROSIS*, janvier 2008.

[3] « Discrimination ». *Wikipédia*, 22 juin 2023. *Wikipedia*, https://fr.wikipedia.org/w/index.php?title=Discrimination&oldid=205483734.

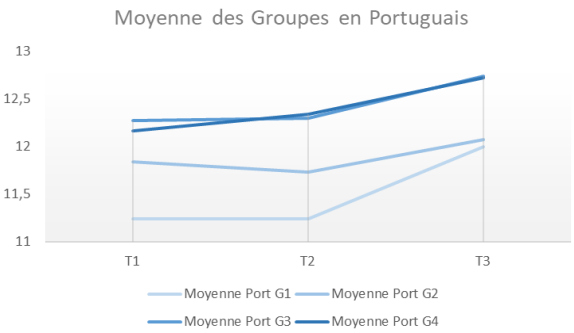
[4] « Égalité entre les filles et les garçons ». *Ministère de l'Education Nationale et de la Jeunesse*, https://www.education.gouv.fr/egalite-entre-les-filles-et-les-garcons-9047. Consulté le 11 juin 2023.

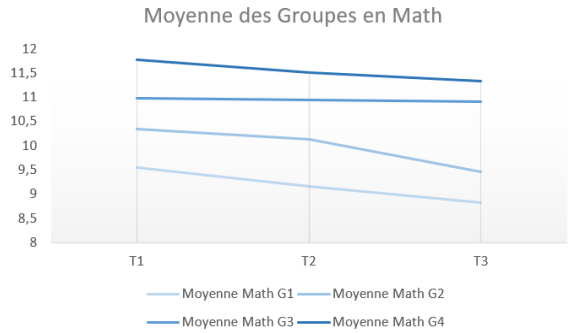
[5] Guilluy, Christophe. « Introduction ». *La France périphérique*, Flammarion, 2015, p. 7-12. *Cairn.info*, https://www.cairn.info/la-france-peripherique--9782081347519-p-7.htm.

[6] Lecocq, Aurélie, et al. « Le score de propension : un guide méthodologique pour les recherches expérimentales et quasi expérimentales en éducation ». *Mesure et évaluation en éducation*, vol. 37, n° 2, 2014, p. 69-100. *www.erudit.org*, https://doi.org/10.7202/1035914ar.

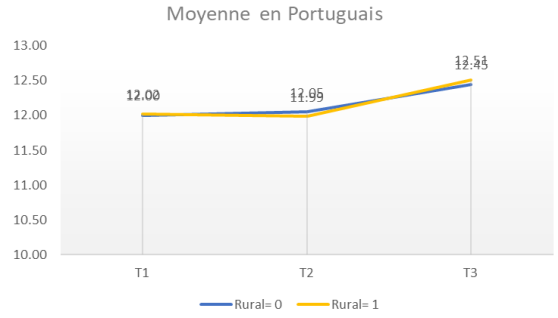
X. Annexes

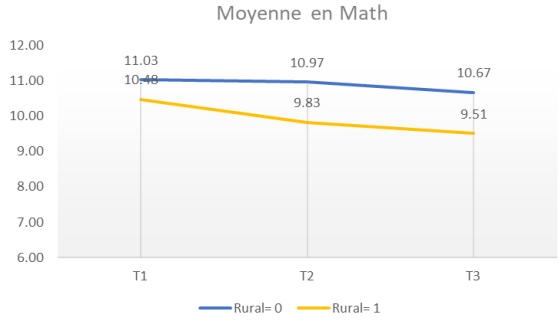
a. Effet de l'éducation des parents des étudiants sur les moyennes en math et portugais

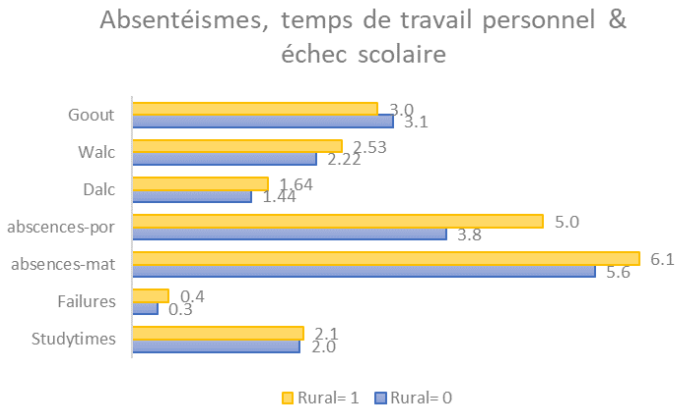




b. Impact de la variable « rural » sur la moyenne scolaire, l'absentéisme, les sorties, la consommation d'alcool, l'échec scolaire et le temps d'étude.







c. Impact de la variable « GP » sur la moyenne scolaire, l'absentéisme, les sorties, la consommation d'alcool, l'échec scolaire et le temps d'étude.