DATA MINING 2

Recherche de facteurs discriminants dans la réussite scolaire

M1 DATA SCIENCES

ANNÉE ACADÉMIQUE 2020-2021

Étudiants :

GOBLET Simon
GEINDREAU Quentin

Enseignant :

DANIEL Christophe

Sommaire

1. Introduction

Qu'est ce qui fait qu'un étudiant réussisse ou non dans l'enseignement secondaire ? Nous aimerions tous connaître la réponse, professeurs comme élèves. Quelqu'un de sérieux, qui travaille régulièrement ou alors quelqu'un qui, malgré des difficultés et une situation familiale difficile, réussi à obtenir la moyenne ?

Ce qui est certain c'est que cette réussite dépend des notes des étudiants. Et beaucoup de paramètres peuvent venir enrayer ou au contraire booster le travail fourni par tous au fur et à mesure de l'année.

Nous allons réaliser une étude dont l'objectif est de déterminer les facteurs susceptibles d'influencer les résultats d'étudiants du secondaire.

Nous commencerons par vous présenter la base de données sur laquelle repose notre étude. Puis, nous commencerons par effectuer une régression linéaire et nous analyserons la pertinence de ce modèle. Enfin, nous chercherons le modèle le plus approprié afin de répondre du mieux possible à notre problématique.

2. Présentation des données

a. Le contexte

La base de données que nous allons étudier à été collectée dans le cadre d'une enquête auprès des 395 élèves de mathématiques dans le secondaire.

Elle contient de nombreuses informations intéressantes sur les questions sociales, de genre et d'étude sur les étudiants des écoles Gabriel Pereira et Mousinho da Silveira.

b. Présentation des variables

Lors de cette étude nous cherchons à expliquer la variables “moyenne” pour chaque étudiant. Notre base de données disposant des notes en première période (G1), en seconde période (G2) et des notes finales (G3), nous avons choisi de calculer la moyenne de ces trois variables afin d’expliquer une variable et non trois.

Pour expliquer notre variable “moyenne”, nous disposons pour chaque élève du secondaire de 30 variables explicatives.

Les notes de chaque élève du secondaire sont caractérisées par les variables :

Les variables quantitatives sont les suivantes :

- age : c’est l’âge de l'élève, ce paramètre est compris entre 15 et 22 ans,
- traveltime : le temps de trajet domicile-école (numérique : 1 correspond à 1 heure),
- studytime : le temps d'étude hebdomadaire de l’élève (numérique : 1 pour 10 heures),
- failures : le nombre d'échecs scolaires antérieurs de l’élève (numérique : 1 ou 2 pour une ou deux années d’échecs, sinon 4),
- absences : le nombre d'absences scolaires de chaque étudiants (numérique : de 0 à 93),

Et voici les variables qualitatives :

- school : l’école de l’élève, cette variable est binaire: “GP” pour Gabriel Pereira ou “MS” pour Mousinho da Silveira,
- sex : le sexe de l’élève, ce paramètre est également biaire : “F” pour le sexe féminin et “M” pour le sexe masculin,
- address : ce paramètre définit le type d'adresse du domicile de l'élève, il est binaire : “U” - urbain ou “R” - rural,
- famsize : la taille de la famille (binaire :”LE3” - inférieure ou égale à 3 ou “GT3”- supérieure à 3),
- Pstatus : le statut de cohabitation des parents (binaire : “T” - vivant ensemble ou “A” - séparés),

- Medu : le niveau d'instruction de la mère (numérique : 0 - aucune, 1 - enseignement primaire (4e année), 2 - de la 5e à la 9e année, 3 - enseignement secondaire ou 4 - enseignement supérieur),
- Fedu : education du père (numérique de 0 à 4 comme pour la variable “Medu”),
- Mjob : la catégorie de l’emploi de la mère (cette variable est nominale : “enseignant", "santé", "services civils” comme par exemple, administration ou police, "à la maison" ou "autre"),
- Fjob : la catégorie de l’emploi du père (nominale et prend les mêmes valeurs que la variable “Mjob”),
- reason : raison du choix de cette école par l’étudiant (nominale : proximité du domicile, réputation de l'école, préférence pour les cours ou autre),
- guardian : le tuteur de l'élève (nominal : 'mère', 'père' ou 'autre'),
- schoolsup : si l’étudiant suit du soutien scolaire supplémentaire (binaire : oui ou non),
- famsup : si l’étudiant dispose de soutien scolaire familial (binaire : oui ou non),
- paid : si l’élève donne des cours supplémentaires rémunérés en mathématiques (binaire : oui ou non),
- activities : si l’étudiant exerce des activités extra-scolaires (binaire : oui ou non),
- nursery : si l’élève a fréquenté l'école maternelle (binaire : oui ou non),
- higher : si l’étudiant souhaite suivre des études supérieures (binaire : oui ou non),
- internet : si l’étudiant a accès à l'internet à la maison (binaire : oui ou non),
- romantic : si l’étudiant a une relation amoureuse (binaire : oui ou non),
- famrel : la qualité des relations familiales de l’étudiant (numérique : de 1 - très mauvaise à 5 - excellente),
- freetime : le temps temps libre après l'école de l’élève (numérique : de 1 - très faible à 5 - très élevé),
- goout : le taux de sorties avec des amis de l’étudiant (numérique : de 1 - très faible à 5 - très élevé),
- Dalc : la consommation d'alcool pendant les journées de cours (numérique : de 1 - très faible à 5 - très élevée),
- Walc : la consommation d'alcool le week-end de l’élève (numérique : de 1 - très faible à 5 - très élevée),

- santé : l'état de santé actuel de l'étudiant (numérique : de 1 - très mauvais à 5 - très bon),

Maintenant ces variables présentées, nous allons effectuer une description statistique de celles-ci. L'ensemble de notre étude sera effectué sur le logiciel R ou bien à l'aide du langage de programmation Python.

c. Analyses descriptives des données

Pour commencer nous allons importer les librairies R suivantes :

- ggplot2 : permet la visualisation des données,
- Hmisc : contient de nombreuses fonctions pour l'analyse de données,
- readr : permet de lire le fichier CSV,
- corrgram : permet de dessiner un corrélogramme,
- car : nécessaire pour l'étude des plus proches voisins,
- lmtest : pour effectuer la régression linéaire,
- FactoExtra, FactoMineR ou encore pls,

Et nous importons notre base que nous appelons stu_data.

On commence par importer la base de données ainsi que de définir notre variable à expliquer "moyenne", sous Python, avec les commandes :

```
data=pd.read_csv(open('student-mat.csv','r'))
data['moyenne']=(data['G1']+data['G2']+data['G3'])/3
data=data.drop(['G1', 'G2','G3'], axis=1)
```

La fonction python suivante nous permet de faire une première description statistique des variables, comme le ferait la fonction summary() sous R ou encore la fonction describe() du package Hmisc.

On obtient les résultats suivants :

	count	mean	std	min	25%	50%	75%	max
age	395.0	16.696203	1.276043	15.000000	16.000000	17.000000	18.000000	22.000000
Medu	395.0	2.749367	1.094735	0.000000	2.000000	3.000000	4.000000	4.000000
Fedu	395.0	2.521519	1.088201	0.000000	2.000000	2.000000	3.000000	4.000000
traveltime	395.0	1.448101	0.697505	1.000000	1.000000	1.000000	2.000000	4.000000
studytime	395.0	2.035443	0.839240	1.000000	1.000000	2.000000	2.000000	4.000000
failures	395.0	0.334177	0.743651	0.000000	0.000000	0.000000	0.000000	3.000000
famrel	395.0	3.944304	0.896659	1.000000	4.000000	4.000000	5.000000	5.000000
freetime	395.0	3.235443	0.998862	1.000000	3.000000	3.000000	4.000000	5.000000
goout	395.0	3.108861	1.113278	1.000000	2.000000	3.000000	4.000000	5.000000
Dalc	395.0	1.481013	0.890741	1.000000	1.000000	1.000000	2.000000	5.000000
Walc	395.0	2.291139	1.287897	1.000000	1.000000	2.000000	3.000000	5.000000
health	395.0	3.554430	1.390303	1.000000	3.000000	4.000000	5.000000	5.000000
absences	395.0	5.708861	8.003096	0.000000	0.000000	4.000000	8.000000	75.000000
moyenne	395.0	10.679325	3.696786	1.333333	8.333333	10.666667	13.333333	19.333333

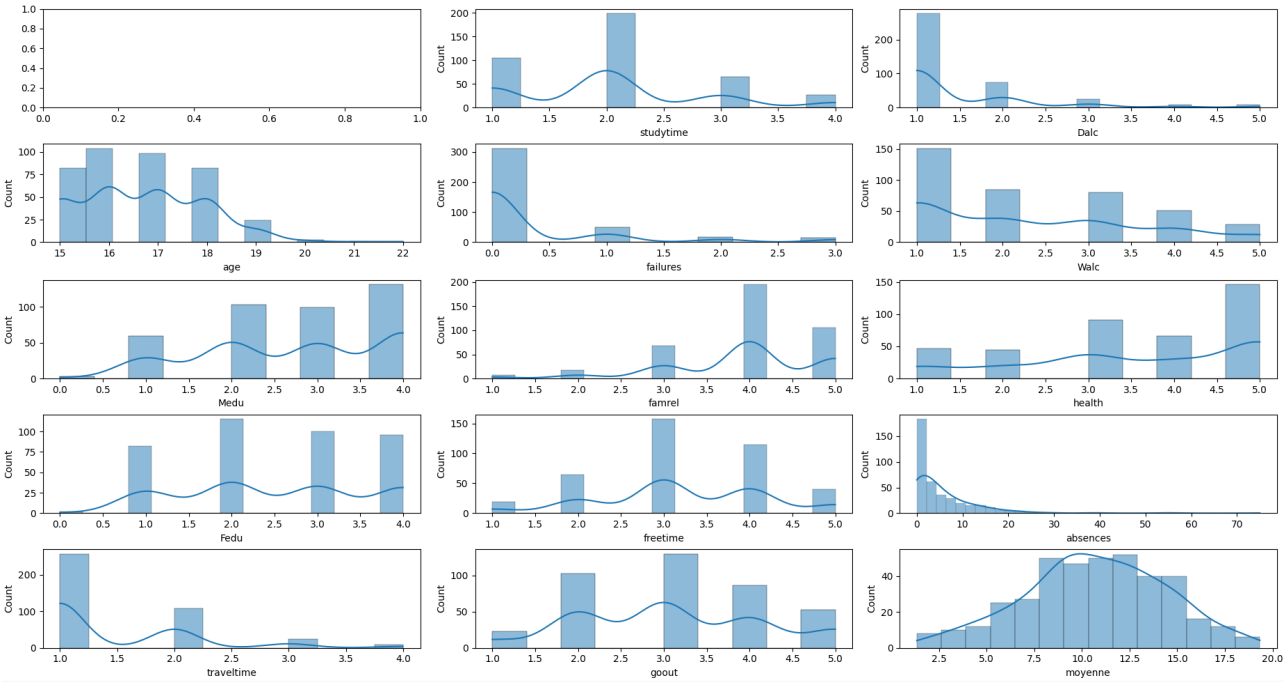
La première remarque que l'on peut faire c'est que pour chaque variable il ne nous manque aucune donnée puisque qu'elles contiennent 395 valeurs, le nombre total d'étudiant interrogés.

Dans la suite de cette partie nous appellerons la variable à expliquer "moy" à la place de moyenne pour ne faire aucun quiproquo lors des analyses descriptives.

On peut observer que notre variable "moy" a un moyenne avoisinant les 10,5. On peut alors penser à une équirépartition des observations de notre base. Cependant en observant le premier et le troisième quartile on voit que la plupart des valeurs sont comprises entre 8 et 13. On peut supposer ici que pour certaines variables nous soyons en présence de valeurs atypiques. Par exemple, pour la variable "absences" , le maximum est atteint en 75, or la moyenne est de 5,7 et le troisième quartile est à 8, ce qui nous laisse penser qu'au moins une valeur est atypique. Il nous faut donc approfondir l'étude de ces variables.

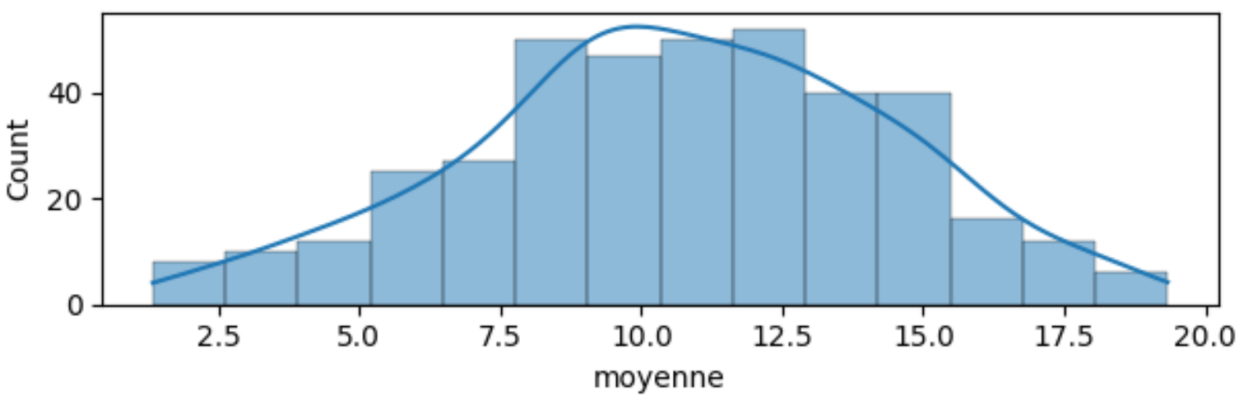
Nous allons donc examiner les données de manière analytique. Une analyse pour que nous puissions voir à quoi ressemblent les variables et comment la variable "moy" se rapporte aux autres individuellement.

Commençons, tout d'abord, à étudier la distribution des variables explicatives à partir d'un histogramme.



D’après ces histogrammes, on peut dire qu’aucune des variables ne semblent symétriques. De ce fait, la distribution ne ressemble visuellement pas à une distribution normale. La variable “absences” semble suivre une distribution exponentielle et il en est de même avec la variable “failures”. Cela signifie que plus la valeur de ces variables sera élevée, moins il y aura d’élèves dans cette situation. Par exemple pour la variable “absences”, on peut constater que la majorité des étudiants n’ont jamais été absents durant l’année. Cependant il y a plusieurs valeurs supérieures à 5 absences et même un élève absent 75 fois.

Concentrons nous maintenant sur l’histogramme de la variable “moyenne” :

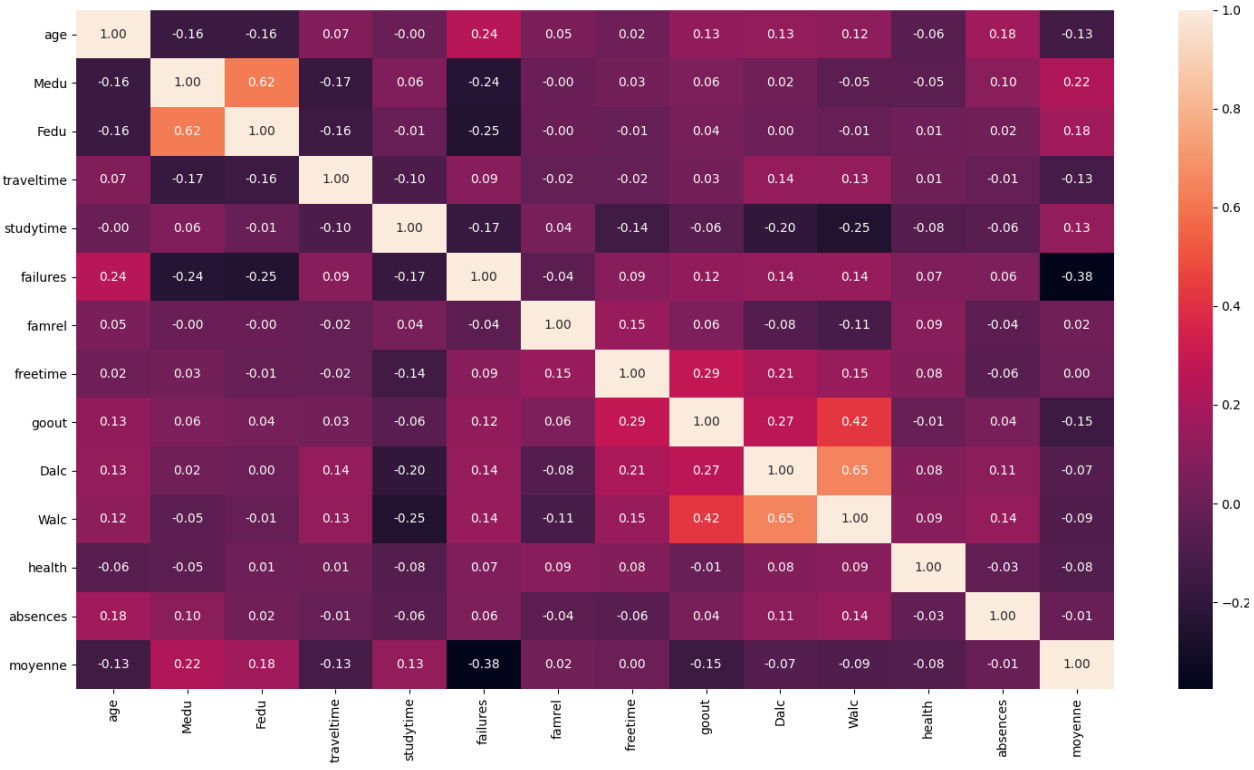


On peut remarquer dans un premier temps que la moyenne n'est pas homogène de 0 à 20.

Autrement dit, il n'y a pas autant de d'élèves ayant eu de très bonnes notes que d'élèves ayant eu des mauvaises notes. Cela peut avoir pour conséquence de rendre difficile à déterminer ce qui caractérise un excellent élève d'un élève moyen.

Cependant, la répartition des moyenne ne se concentre pas sur quelques notes. Il n'est donc pas nécessaire ici de distinguer plusieurs catégories d'étudiants en fonction de leur moyenne.

Nous avons bien étudié chaque variable séparément, voyons à présent comment celles-ci sont liées. Une méthode pour cela consiste à étudier leur corrélation, ainsi que leur répartition. Dans notre sortie ci-dessous, la variable "moyenne" affichera les valeurs des corrélations entre variables.



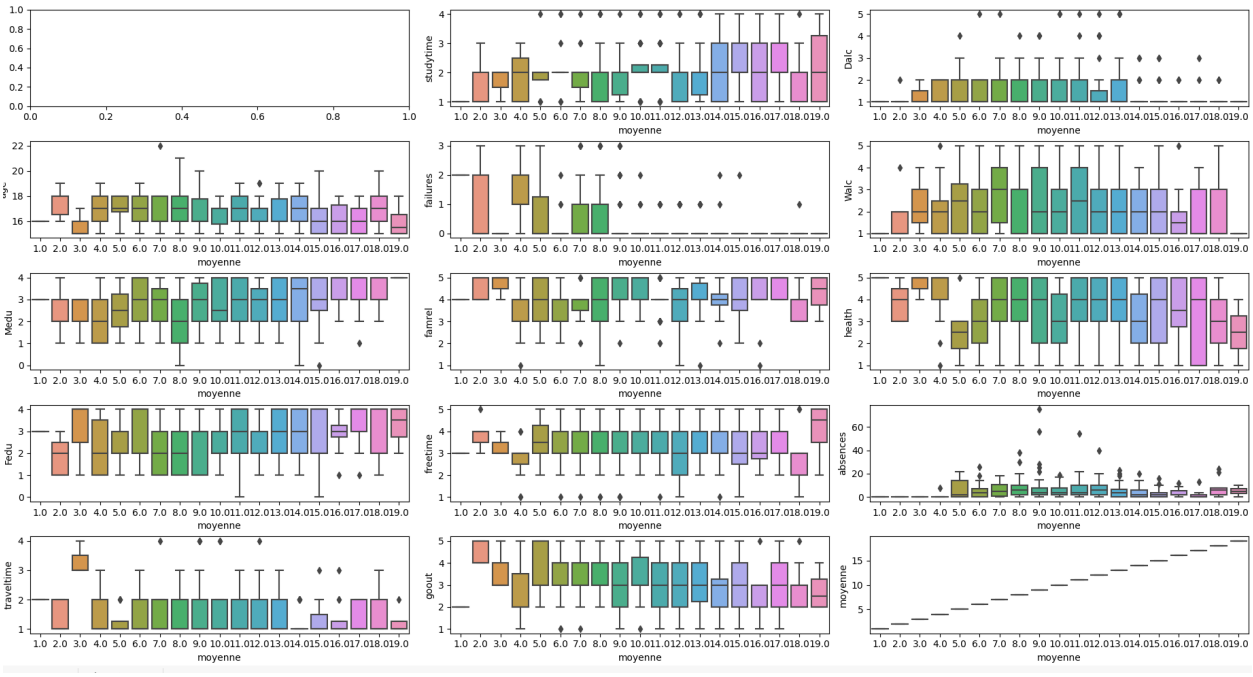
Sur cette première figure, qui représente les corrélations entre variables, nous pouvons constater qu'aucunes variables ne semble fortement corrélé (aucune valeur n'est supérieure à 0,70). On constate cependant, que certaines paires de variables sont corrélées de façon raisonnable.

Par exemple, le fait qu'un étudiant sorte le week-end est décemment corrélé (corrélation légèrement supérieur à 0,4) avec sa consommation d'alcool durant le week-end.

On peut également voir une corrélation non négligeable (corrélation de 0,65) entre les variables Dalc et Walc. Ce qui signifie que la consommation d'alcool d'un étudiant durant la semaine dépend de sa consommation le week-end.

Nous allons maintenant regarder les diagrammes en boîte de nos variables explicatives puisqu'ils peuvent être plus faciles à interpréter visuellement que les valeurs de corrélation.

De plus, il existe une forte concentration d'observation entre 8 et 15 pour la moyenne. De ce fait, les valeurs de corrélation peuvent être trompeuses et davantage basées sur le comportement des valeurs moyennes plutôt que sur les autres observations.



Ces diagrammes nous indiquent les tendances de chaque composante quantitative en fonction de la moyenne de l'élève.

Par exemple, il semblerait que la moyenne d'un étudiant élevé soit associé à un temps d'étude également élevé. De plus, pour les absences des élèves, on peut observer qu'un

nombre d'absences important pour un élève soit plus fréquent avec une moyenne assez basse (entre 5 et 9).

Pour les variables freetime et goout, la répartition de toutes les boîtes à moustaches semble homogène. De ce fait, on remarque que la proportion de temps libre prise par les étudiant impact faiblement leurs moyenne. Il y a même davantage de temps libre pris par les élèves ayant eu 19.

Concernant la variable failures, on peut observer que les étudiants ayant redoublé restent dans la moyenne basse des élèves (entre 1 et 8).

Enfin pour la variable Dalc, il semble que les étudiants ayant les meilleur moyenne (entre 14 et 19) ne consomment pas d'alcool durant une journée de cours.

Nous avons donc une première approche de nos variables et de l'impact qu'elles ont sur la variable moyenne. Il est temps maintenant d'approfondir notre étude. Dans un premier temps nous allons effectuer une régression linéaire multiple et voir si ce modèle peut répondre à notre problématique.

3. Régression Linéaire Multiple

a. Détails de la régression linéaire multiple

Le but de cette régression est de minimiser la variance de la variable qualité, c'est la méthode des MCO. Nous allons donc choisir les composantes, et retirer celles qui ne sont pas significatives, dans le but de résoudre cette optimisation.

Nous effectuons une méthode backward pour sélectionner les variables. Dans cette méthode, le but est donc de minimiser l'AIC. Pour commencer, nous effectuons une régression linéaire avec l'ensemble des variables.

```
> extractAIC(ds.lm)
[1]  40.0000 964.6018
```

Ici, l'AIC à minimiser est de 964.6018.

```
Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     13.026412   3.525493   3.695 0.000255 ***
schoolMS         0.329757   0.622793   0.529 0.596803
sexM             1.038900   0.393415   2.641 0.008638 **
age             -0.214090   0.170901  -1.253 0.211134
addressU         0.379715   0.459576   0.826 0.409230
famsizeLE3       0.581104   0.384140   1.513 0.131235
PstatusT        -0.132412   0.569551  -0.232 0.816295
Medu             0.297769   0.254265   1.171 0.242345
Fedu             0.013837   0.218427   0.063 0.949526
Mjobhealth       0.979016   0.879771   1.113 0.266544
Mjobother       -0.479789   0.561106  -0.855 0.393085
Mjobservices     0.557146   0.627727   0.888 0.375378
Mjobteacher     -1.083875   0.816847  -1.327 0.185394
Fjobhealth      -0.026356   1.131366  -0.023 0.981427
Fjobother       -0.696878   0.804913  -0.866 0.387195
Fjobservices    -0.471243   0.831610  -0.567 0.571300
Fjobteacher      1.260573   1.020097   1.236 0.217373
reasonhome       0.170654   0.435724   0.392 0.695547
reasonother      0.374018   0.643253   0.581 0.561308
reasonreputation 0.491654   0.453633   1.084 0.279183
guardianmother  -0.007161   0.429267  -0.017 0.986700
guardianother    0.740753   0.786362   0.942 0.346833
traveltime      -0.204244   0.266693  -0.766 0.444280
studytime        0.572795   0.226322   2.531 0.011809 *
failures        -1.470681   0.261925  -5.615 3.97e-08 ***
schoolsupyes    -1.656124   0.524730  -3.156 0.001735 **
famsupyes       -0.910809   0.376626  -2.418 0.016094 *
paidyes          0.177769   0.375885   0.473 0.636551
activitiesyes   -0.118201   0.350074  -0.338 0.735829
nurseryyes      -0.035372   0.432190  -0.082 0.934817
higheryes        1.160534   0.847994   1.369 0.172001
internetyes      0.458134   0.487457   0.940 0.347935
romanticyes     -0.707732   0.369198  -1.917 0.056047 .
famrel           0.040432   0.193490   0.209 0.834597
freetime         0.257924   0.186746   1.381 0.168102
goout           -0.519451   0.176638  -2.941 0.003489 **
Dalc            -0.138005   0.260327  -0.530 0.596360
Walc             0.111173   0.195130   0.570 0.569217
health          -0.187243   0.126681  -1.478 0.140279
absences         0.025674   0.022790   1.127 0.260696
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.232 on 355 degrees of freedom
Multiple R-squared:  0.3113,    Adjusted R-squared:  0.2356
F-statistic: 4.114 on 39 and 355 DF,  p-value: 3.82e-13
```

On remarque qu’au seuil de 5%, seule la variable failures est significative. Nous avons donc 1 variable à expliquer sur 30 qui ne sont pas significatives. De plus, le R2 Ajustée est de 0,2356, ce qui signifie que le modèle explique 23,56% de la variance de la variable qualité.

Nous allons donc retirer les variables non significatives et effectuer une nouvelle régression linéaire multiple sur ce nouveau modèle.

```
Call:
lm(formula = moyenne ~ sex + studytime + failures + schoolsup +
    goout, data = ds)

Residuals:
    Min      1Q  Median      3Q     Max
-9.0202 -1.9359  0.0746  2.2257  9.6719

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   11.1342     0.7299  15.255  < 2e-16 ***
sexM           1.0606     0.3568   2.972  0.00314 **
studytime      0.5082     0.2134   2.381  0.01773 *
failures      -1.7263     0.2310  -7.473 5.22e-13 ***
schoolsupyes  -1.3964     0.5055  -2.762  0.00601 **
goout         -0.3970     0.1526  -2.602  0.00963 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.335 on 389 degrees of freedom
Multiple R-squared:  0.1964,    Adjusted R-squared:  0.1861
F-statistic: 19.02 on 5 and 389 DF,  p-value: < 2.2e-16

[1]   6.0000 957.5037
```

Avec ce modèle, on remarque que l'AIC est de 957.3777. On a donc une diminution de l'AIC. De plus, le R2 Ajustée a diminué et est de 0,1884 , ce qui signifie que le modèle explique 18,84% de la variance de la variable qualité, ce qui est moins que le modèle précédent.

Cependant, le but est de minimiser le critère de l'AIC nous conserverons donc le second modèle.

Maintenant nous allons approfondir notre étude afin d'évaluer la pertinence du modèle.

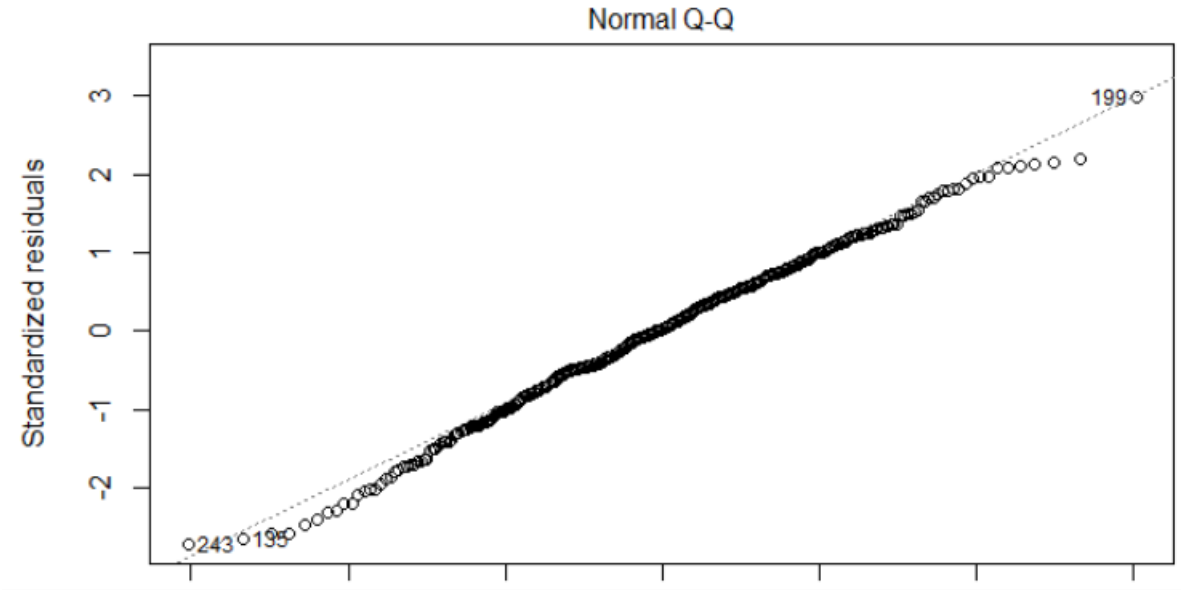
b. Evaluation de la pertinence de notre modèle

Afin d'examiner la pertinence du modèle, nous allons étudier la validité de différentes hypothèses (normalité, homoscédasticité, adéquation...). Pour cela, nous effectuons les tests correspondants, avec l'ajout de certains graphiques afin de faciliter l'interprétation.

Pour commencer, étudions l'hypothèse de normalité à partir du test de shapiro.

```
        Shapiro-Wilk normality test

data:  resid(ds.lm2)
W = 0.99507, p-value = 0.2415
```

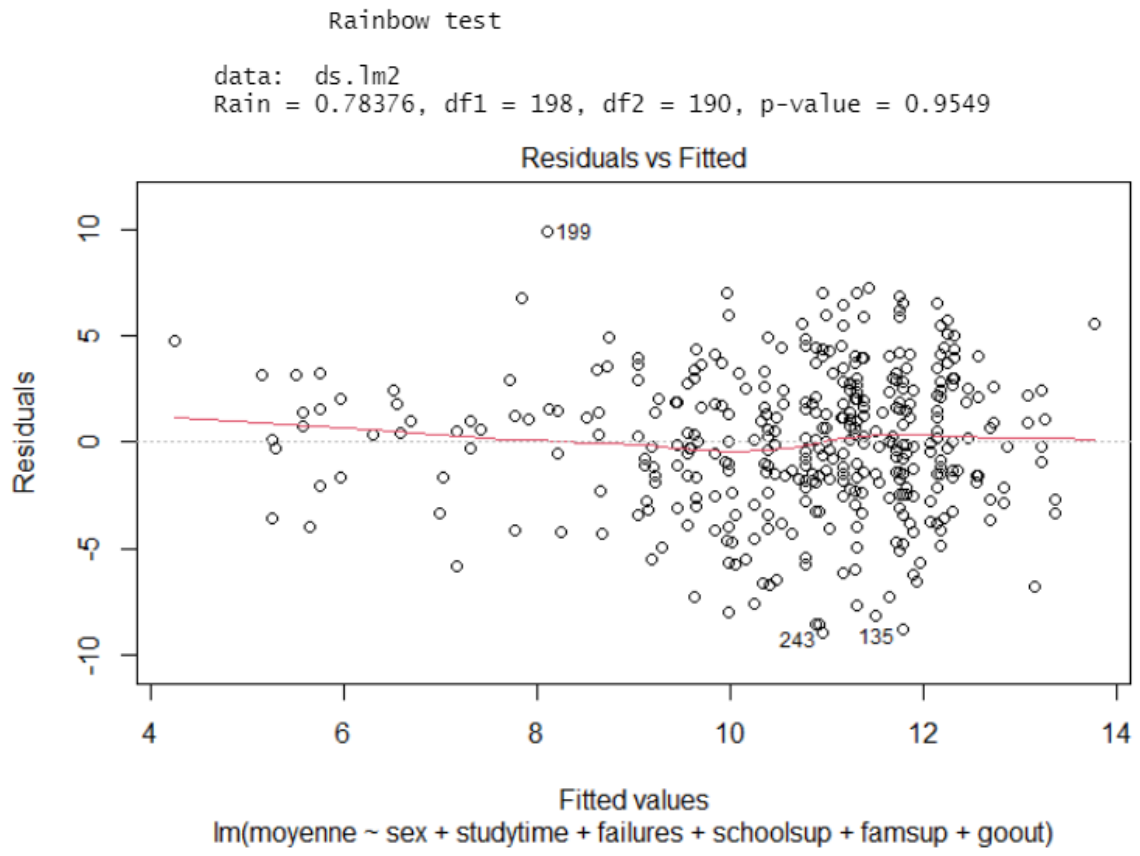
On remarque que la p-value n'est pas



significative au seuil de 5%. On en déduit que notre échantillon suit une loi normale.

Graphiquement, nous avons effectué un QQ plot. Cet outil nous permet de comparer la distribution de notre échantillon avec celle d'une loi gaussienne réduite. Ici, on remarque que presque tous les points sont alignés sur la première bissectrice, ce qui confirme que notre distribution suit une loi de distribution gaussienne normalisée.

Vérifions maintenant les hypothèses d'adéquation à partir du test de Rainbow.



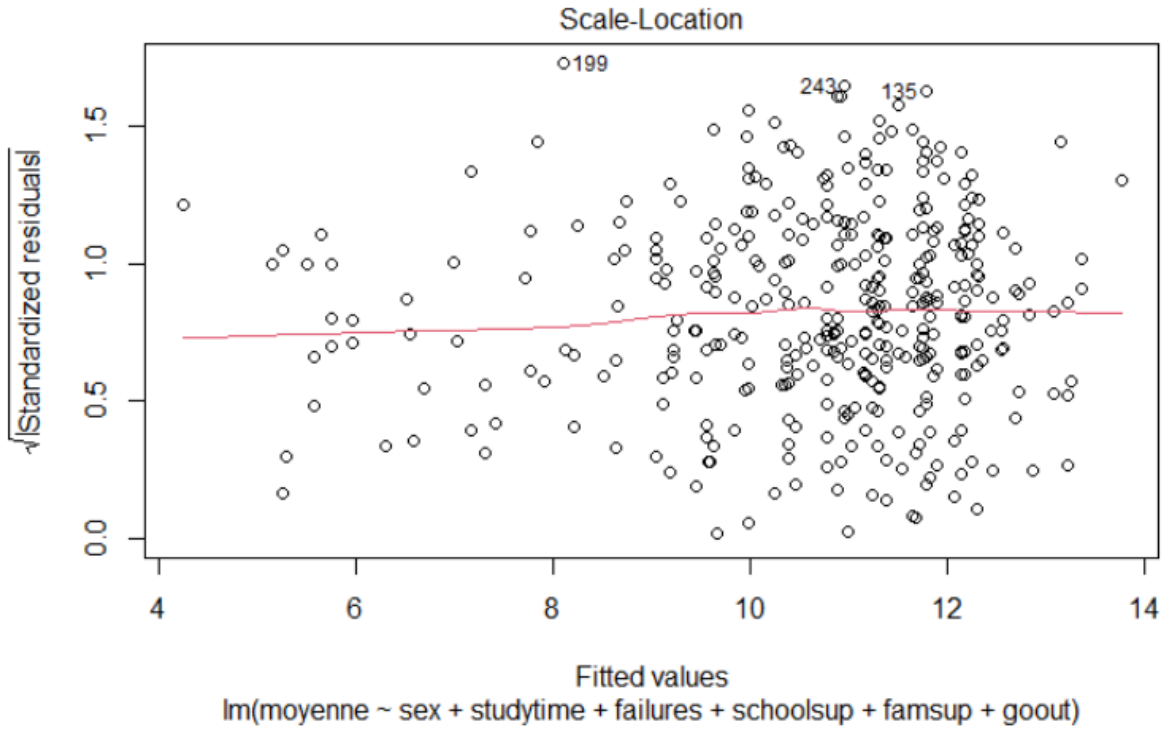
Dans un premier temps, étudions le test de Rainbow sur l'adéquation des résidus. On constate que la p-value est très proche de 1, on en conclut donc qu'il y a adéquation du modèle.

Nous avons effectué un graphique afin de mieux visualiser cette adéquation. En effet, la ligne rouge doit être horizontale s'il y a une relation entre la moyenne et les autres

variables. Or cette ligne l'est quasiment. On en conclut que le modèle de régression semble adapté.

Regardons maintenant les hypothèses d'adéquation à partir du test de Golfeld-Quandt.

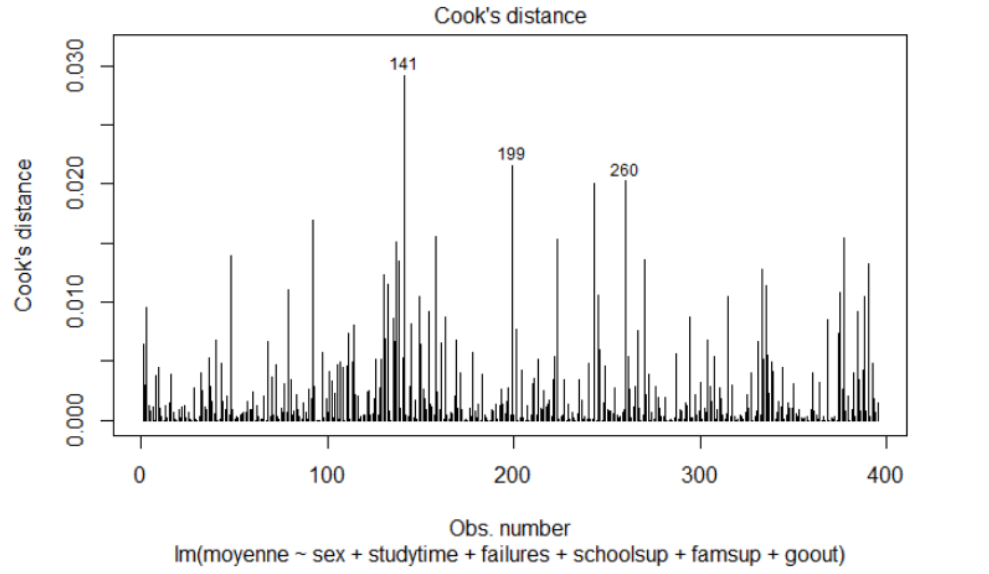
```
        Goldfeld-Quandt test

data:  ds.lm2
GQ = 1.0854, df1 = 191, df2 = 190, p-value = 0.2861
alternative hypothesis: variance increases from segment 1 to 2
```

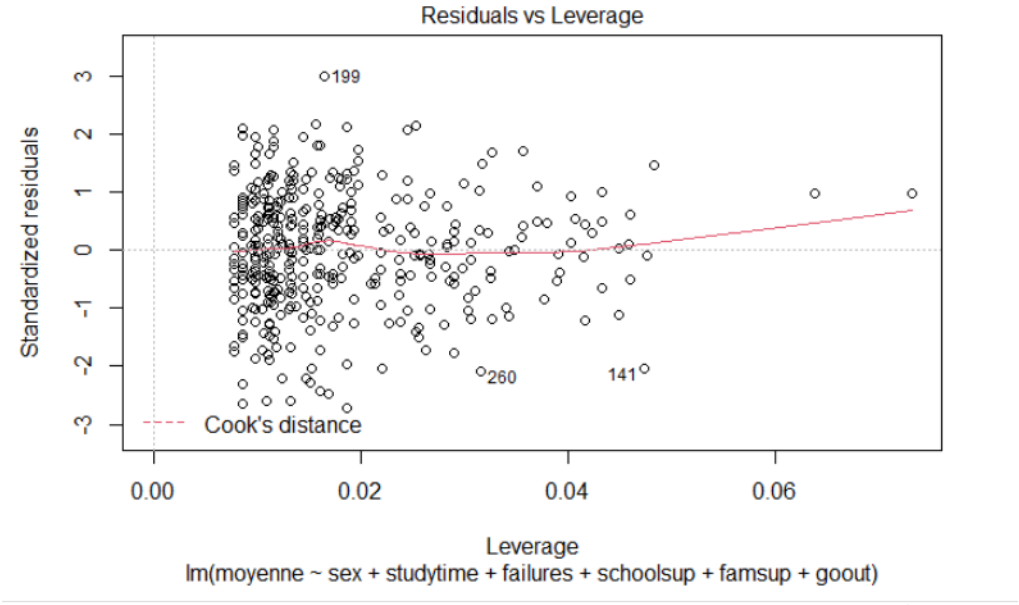


Comment interpréter ce test ? Ce test nous permet de vérifier objectivement l'homogénéité des résidus (et donc leurs l'hétéroscédasticité). Avec le test de Goldfeld-Quandt, on remarque que la p-value est supérieure à 0,05, ce qui signifie qu'il y a homogénéité des résidus au seuil de 5%. Pour faciliter cette interprétation nous avons effectué une représentation graphique qui met en relation les racines carrées des résidus en fonction des valeurs théoriques de la variable qualité prédite par l'équation de régression.

L'homogénéité est à rejeter si la courbe rouge n'est pas horizontale, ce qui n'est pas le cas ici. Donc il y a une distribution homogène des résidus, et de ce fait la variance des résidus est constante.

Nous allons maintenant rechercher la présence de valeurs aberrantes. Pour cela nous allons effectuer uniquement une interprétation graphique à l'aide de la distance de Cook.





L'objectif ici est d'évaluer les points qui ont une grande influence sur la régression, afin de les écarter s'il s'agit de points potentiellement aberrants. La distance de Cook nous permet d'évaluer les points qui auront une (trop) grande influence sur le modèle de régression. À partir du premier graphique représentant cette distance, on remarque que sur 395 observations, seules 3 observations semblent aberrantes : les observations n°141, n°199 et n°260. Ceci est également confirmé dans notre second graphique.

On en conclut que seules 3 observations semblent aberrantes, ce qui est négligeable au vu du nombre d'observations totales.

Enfin, nous allons analyser la multicolinéarité des variables à l'aide la fonction vif()

```
      sex studytime  failures schoolsup    famsup     goout
1.139047  1.147524  1.046479  1.028131  1.043888  1.022316
```

Ici, les vif de chaque variable estiment de combien la variance d'un coefficient est "augmentée" en raison d'une relation linéaire avec d'autres prédicteurs.

Par exemple, le vif de la variable sex est environ de 1,13, cela nous indique que la variance de ce coefficient particulier est supérieure à environ 13% à celle que l'on aurait dû observer si ce facteur n'est absolument pas corrélé aux autres prédicteurs. On constate que tous les vif sont très proches de 1, on en conclut qu'il n'y a donc pas de problème potentiel de colinéarité à explorer.

c. Conclusion de la régression linéaire multiple

À la suite de tous ces tests, nous pouvons tirer plusieurs conclusions. Tout d'abord, le test de Shapiro nous indique que notre échantillon suit une loi normale. Ceci nous informe que notre modèle de régression linéaire multiple est adapté à notre échantillon. Ceci nous a ensuite été confirmé par le test de Rainbow, nous indiquant l'existence d'adéquation avec le modèle. De plus, le test de Goldfeld-Quandt nous précise que la distribution des résidus est homogène, et donc que la variance de nos résidus est constante.

Nous avons ensuite poursuivi l'étude de ce modèle à l'aide de la distance de Cook. Nous avons découvert seulement 3 observations qui semblaient aberrantes, ce que nous avons considéré négligeable au vu de la quantité d'observations contenues dans notre échantillon. Enfin, la fonction vif() nous a informé que notre modèle ne présentait pas de problème d'une éventuelle colinéarité.

À partir de toutes ces analyses, on en conclut que le modèle de régression linéaire multiple est un modèle pertinent pour notre base de données, mais il n'est peut-être pas le plus adapté pour répondre à notre problématique.

Nous allons donc poursuivre notre étude à partir de modèles qui semblent également pertinents mais plus adaptés à notre problématique.

4. Etude en composantes principales

L'objectif de cette partie va être de simplifier notre base de données afin d'en tirer la meilleure interprétation. Nous avons ici choisi d'appliquer une Régression en Composante Principale (PCR) qui nous semble la plus adaptée à notre base de données. En effet, notre nombre de variables étant très largement inférieur au nombre d'observations, appliquer la méthode de Régression des Moindres Carrés Partiels (PLS) n'aurait pas été pertinent (cette analyse en PLS sera en annexe).

Pour appliquer la méthode de Régression en Composante Principale nous allons procéder en deux étapes.

Tout d'abord, nous allons effectuer une Analyse en Composantes Principales (ACP) puis nous effectuerons une Régression Multiple sur ces composantes principales.

a. Analyse en composantes principales

Cette analyse va nous permettre de synthétiser les informations importantes contenues dans notre base de données, en seulement quelques variables : les composantes principales.

Ces composantes principales (ou axes principaux), correspondent à une combinaison linéaire de notre variable. L' information contenue dans notre base données

correspond à la variance (aussi appelé inertie totale). L'objectif de notre ACP va donc être d'identifier les directions le long desquelles la variation des données est maximale.

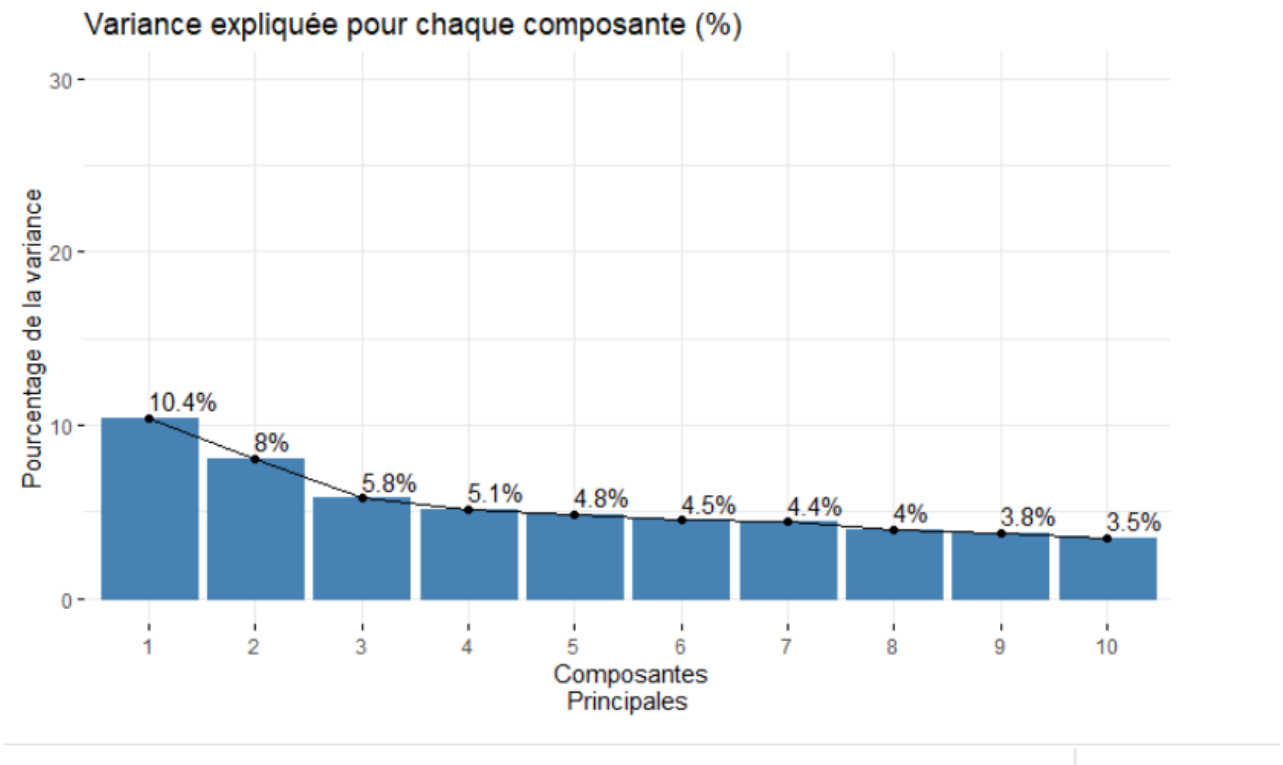
Pour commencer, nous allons étudier les valeurs propres de cette ACP car elles mesurent la quantité de variance à expliquer. Cela va nous permettre de déterminer le nombre de composantes principales nécessaires pour notre étude.

Pour cette étude il faut que toutes les variables soient quantitatives. De ce fait, nous avons remplacé chaque variable qualitative par un entier qui la désigne et de cette façon nous pouvons effectuer l'ACP de notre modèle.

	eigenvalue	percentage of variance	cumulative percentage of variance
comp 1	3.2223786	10.3947698	10.39477
comp 2	2.4934635	8.0434306	18.43820
comp 3	1.7874577	5.7659925	24.20419
comp 4	1.5852889	5.1138352	29.31803
comp 5	1.5023536	4.8463020	34.16433
comp 6	1.3951857	4.5005991	38.66493
comp 7	1.3701335	4.4197855	43.08471
comp 8	1.2280340	3.9614000	47.04611
comp 9	1.1816103	3.8116460	50.85776
comp 10	1.0902417	3.5169087	54.37467
comp 11	1.0463981	3.3754778	57.75015
comp 12	1.0107102	3.2603553	61.01050
comp 13	0.9581046	3.0906600	64.10116
comp 14	0.9375598	3.0243866	67.12555
comp 15	0.8710351	2.8097905	69.93534
comp 16	0.8200980	2.6454775	72.58082
comp 17	0.8000760	2.5808902	75.16171
comp 18	0.7669566	2.4740535	77.63576
comp 19	0.7407014	2.3893592	80.02512
comp 20	0.7109802	2.2934845	82.31860
comp 21	0.6983186	2.2526408	84.57125
comp 22	0.6290672	2.0292490	86.60049
comp 23	0.5958058	1.9219541	88.52245
comp 24	0.5912160	1.9071483	90.42960
comp 25	0.5593081	1.8042196	92.23382
comp 26	0.5139704	1.6579689	93.89179
comp 27	0.4872674	1.5718303	95.46362
comp 28	0.4391406	1.4165825	96.88020
comp 29	0.3919240	1.2642710	98.14447
comp 30	0.3016789	0.9731578	99.11763
comp 31	0.2735357	0.8823734	100.00000

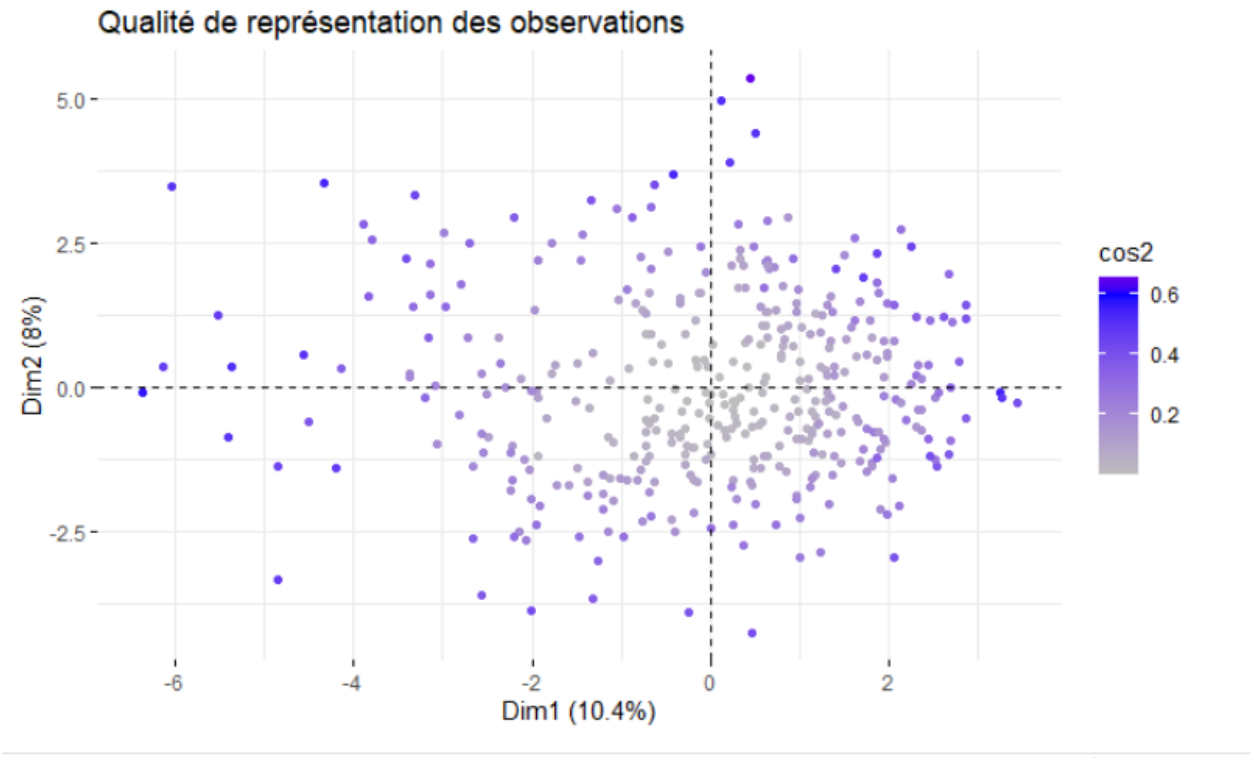
Nous observons en sortie, trois coefficients pour chaque composante. La valeur propre de la composante, sachant que la somme des valeurs propres vaut 10. Puis le coefficient de variance pour chaque composante. Par exemple, la première composante explique environ 10,39% de la variance. Enfin, la dernière colonne nous indique la variance cumulée pour chaque composante. On observe que prendre les 3 premières composantes explique environ 24,20% de la variance, ce qui nous semble acceptable en raison du nombre de composantes.

Une interprétation graphique à partir de la méthode du « coude » peut nous permettre de valider ce choix, puisqu'on peut voir une "cassure" au niveau de la troisième composante principale.



Le graphique affiche le pourcentage de variance expliquée pour chaque composante principale. On remarque qu'à partir de la troisième composante, le pourcentage de la variance est inférieur à 5,8%. Ceci confirme donc bien notre choix.

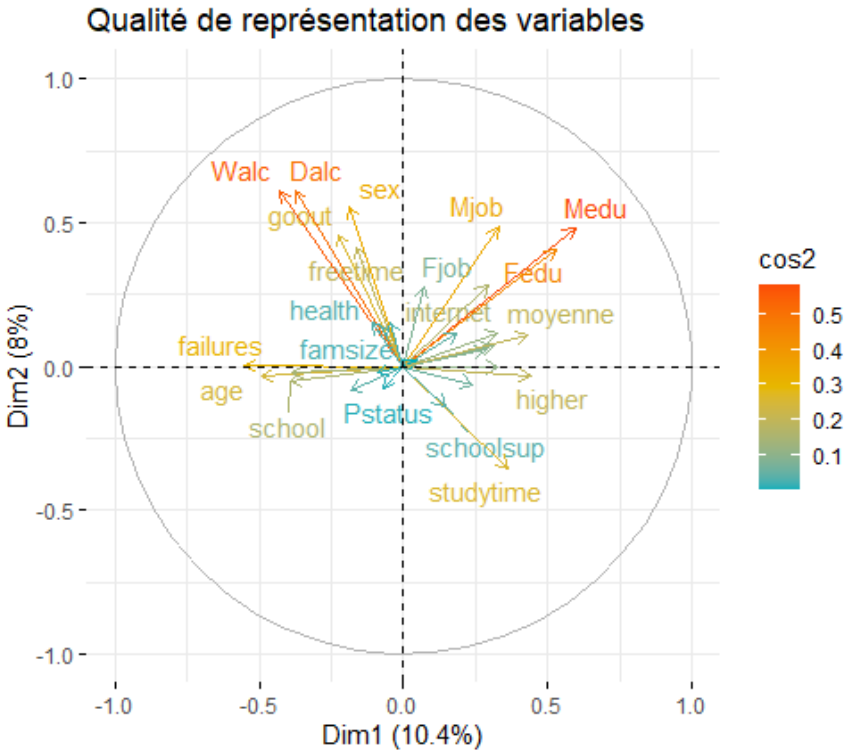
Maintenant nous allons visualiser graphiquement la qualité de représentation de nos observations sur ces composantes. Il est possible de visualiser cette qualité à partir du cos2 de chaque étudiant. Nous avons donc effectué une visualisation de ces observations uniquement à partir des deux premières composantes.



Un cos2 élevé nous indique une bonne représentation de nos observations sur les composantes principales.Plus simplement, plus l'observation est bleue, mieux elle est représentée par les deux premières composantes.

Une observation se situant donc proche du centre du cercle ne sera pas bien représentée. On remarque que plusieurs étudiants sont situés dans cette zone, et sont donc en gris. Cela confirme notre nécessité de prendre plus de deux composantes pour pouvoir parfaitement représenter l'ensemble de nos observations.

Passons ensuite à la qualité de représentation de nos variables explicatives sur ces deux composantes à l'aide de leur cos2.

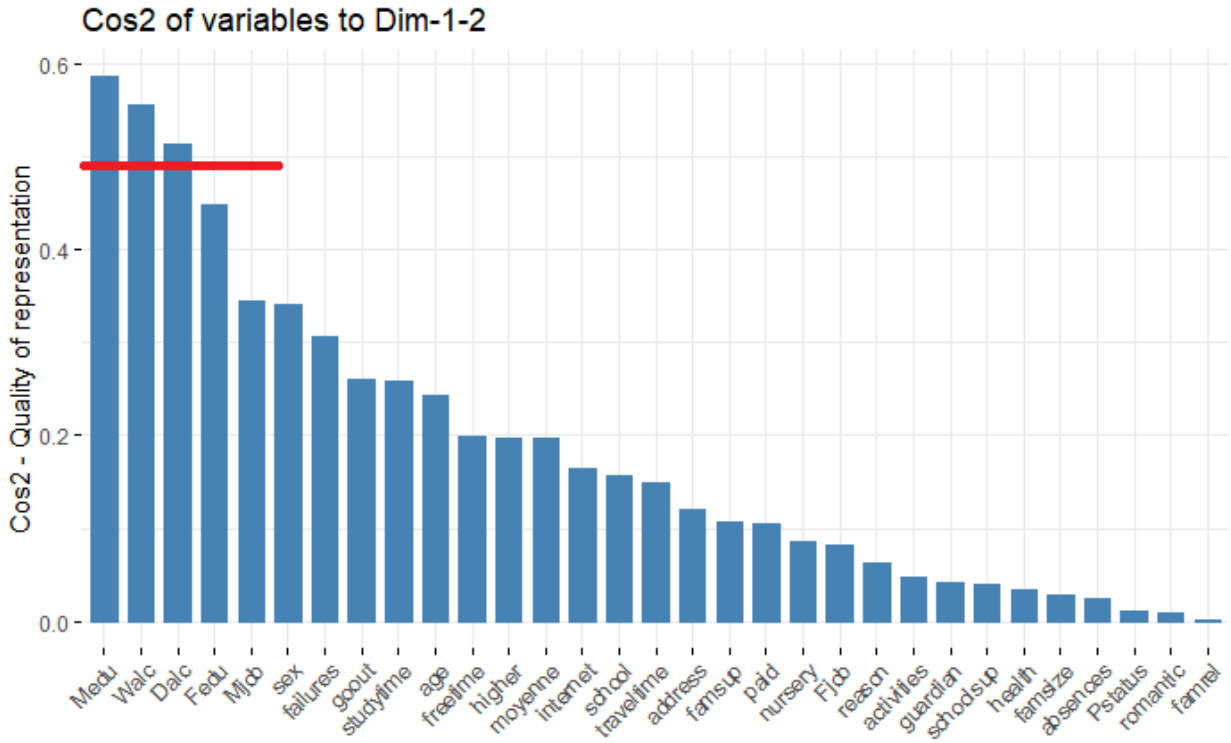


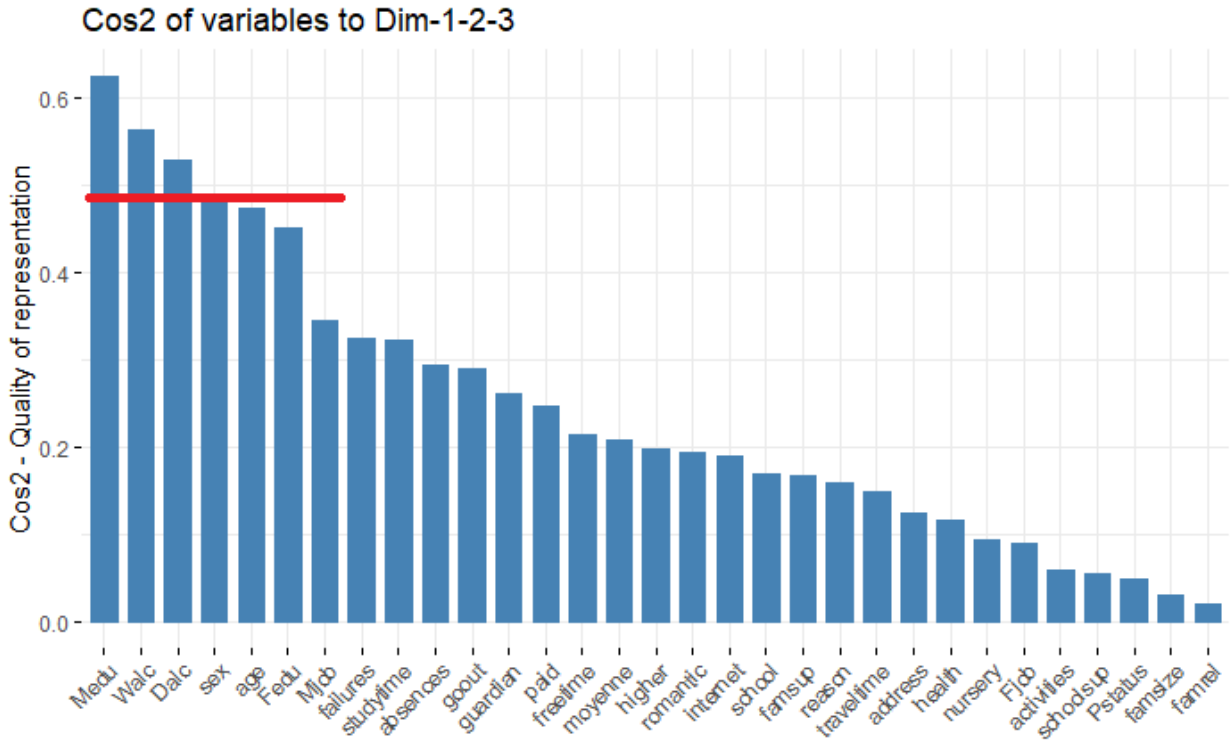
Ce graphique est un cercle de corrélation, il nous indique la qualité de représentation de chaque variable en fonction des deux premières composantes. Plus une variable est proche de ce cercle, donc plus elle est orange, plus elle est corrélée. Une variable orange sera donc parfaitement représentée, alors qu'une variable bleue ne sera pas représentée par ces composantes.

On remarque que les variables Walc, Dalc, Fedu et Medu semblent bien représentées, et pas particulièrement par une des deux composantes en particulier.

Les composantes 1 représentent principalement la variable sex, tandis que les composantes 2 vont principalement représenter les variables failures, higher,age, moyenne et school. Pour ce qui est des variables restantes, celles-ci n'ont pas une représentation satisfaisante dans ces deux composantes. Ce qui nous conforte dans la nécessité de retenir plus de composantes, et en l'occurrence trois.

Comparons maintenant le cos2 de chaque variable lorsque nous retenons deux composantes et lorsque nous en retenons trois.

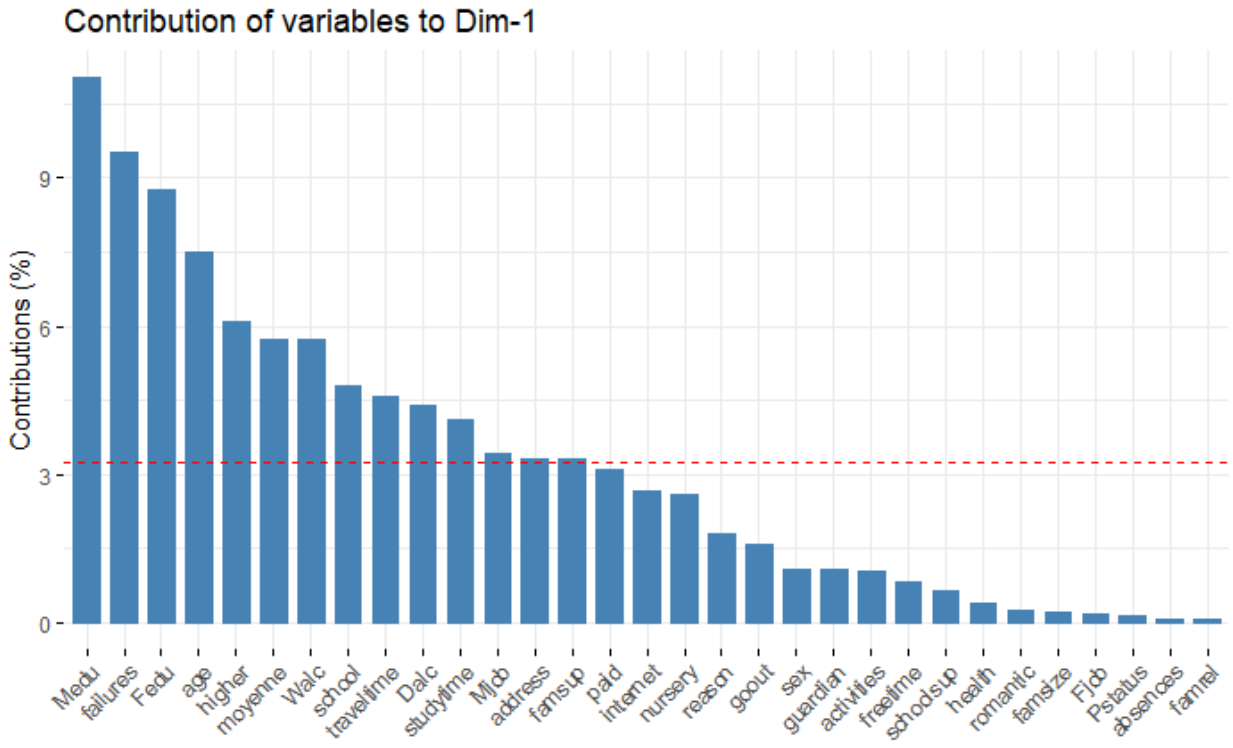

Cos2 of variables to Dim-1-2
Cos2 - Quality of representation
0.6
0.4
0.2
0.0
Medu
Walc
Dalc
Fedu
Mjob
sex
failures
goout
studytime
age
freetime
higher
moyenne
internet
school
traveltime
address
famsup
paid
nursery
Fjob
reason
activities
guardian
schoolsup
health
famsize
absences
Pstatus
romantic
famrel


Cos2 of variables to Dim-1-2-3
Cos2 - Quality of representation
0.6
0.4
0.2
0.0
Medu
Walc
Dalc
sex
age
Fedu
Mjob
failures
studytime
absences
goout
guardian
paid
freetime
moyenne
higher
romantic
internet
school
famsup
reason
traveltime
address
health
nursery
Fjob
activities
schoolsup
Pstatus
famsize
famrel

Ces deux histogrammes nous montrent qu'avec deux composantes, il y a trois variables bien représentées (corrélation supérieure à 0,48 représentée par la ligne rouge) : Medu, Walc et Dalc.

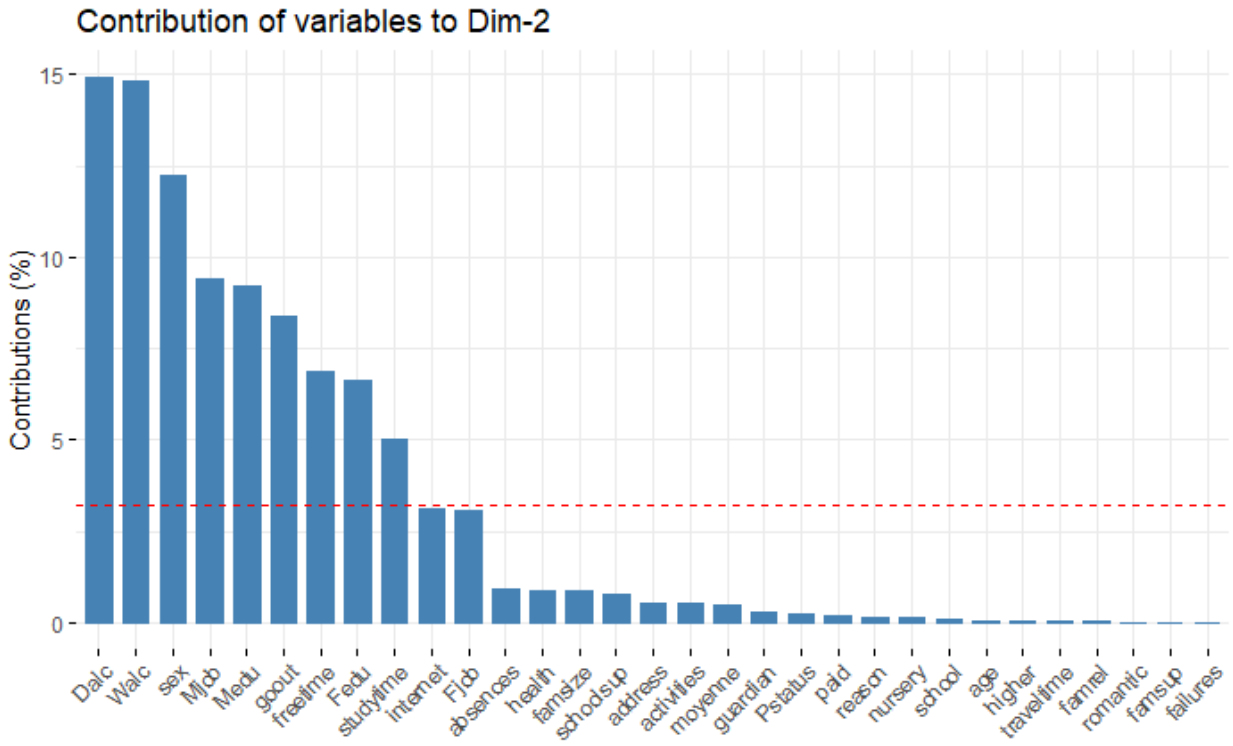
En retenant trois composantes, les variables sex, age et Fedu sont représentées correctement également.

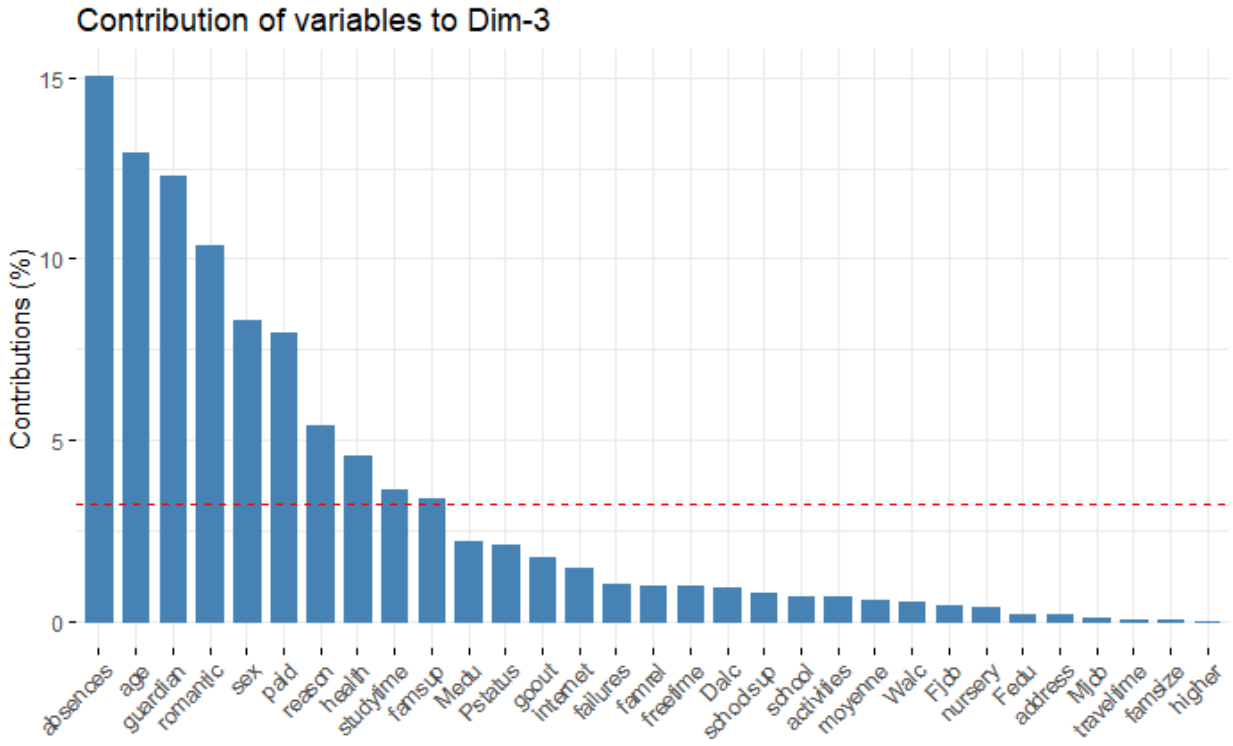
Analysons ensuite la contribution des variables pour chaque composante.



Ce graphique nous montre que quatorze variables ont une forte contribution pour l'axe 1, que l'on appellera PC1: Medu, failures, Fedu, age, higher, moyenne, walc, school, traveltime, Dalc, studytime, Mjob, address et famsup.

Sur le deuxième histogramme, il n'y a cette fois-ci que neuf variables qui ont une forte contribution sur l'axe 2 (PC2) : Dalc, Walc, sex, Mjob, Medu, goout, freetime, Fedu et studytime.





Enfin ce cette dernière représentation les variables qui ont une forte contribution pour l'axe 3 (PC3) sont : absences, age, guardian, romantic, sex, paid, reason, health, studytime et

famsup. L'ensemble de nos composantes étant déterminées, nous allons maintenant appliquer une régression multiple.

b. Régression sur les composantes principales

L'objectif de la régression (ou PCR) est de calculer les composantes principales, puis d'utiliser certaines de ces composantes comme prédicteurs dans un modèle de régression linéaire ajusté en utilisant la procédure des moindres carrés ordinaire (MCO).

Pour commencer, nous avons scindé notre base de données, où Y représente notre variable à expliquer moyenne, et X représente nos variables explicatives.

On obtient les résultats suivants :

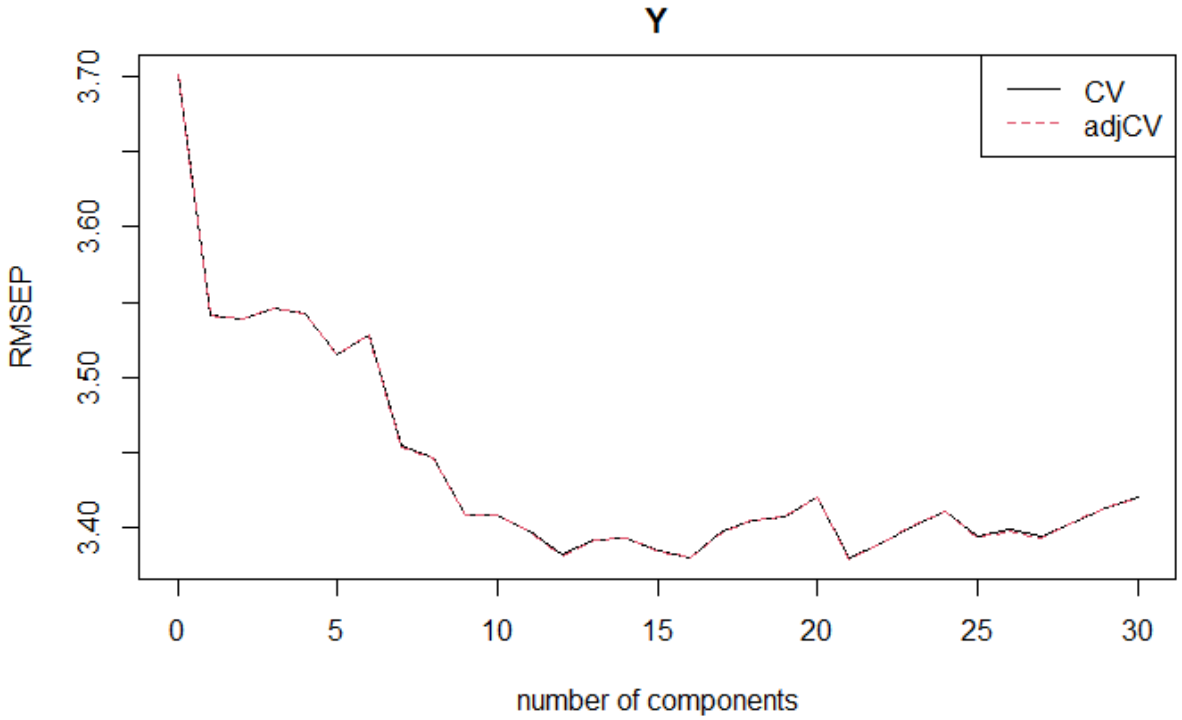
```
Principal component regression , fitted with the singular value decomposition algorithm.
Cross-validated using 395 leave-one-out segments.
Call:
pcr(formula = Y ~ X, ncomp = 30, data = ds, scale = TRUE, validation = "LOO")
Data:   X dimension: 395 30
        Y dimension: 395 1
Fit method: svdpc
Number of components considered: 30

VALIDATION: RMSEP
Cross-validated using 395 leave-one-out segments.
       (Intercept)  1 comps  2 comps  3 comps  4 comps  5 comps  6 comps  7 comps  8 comps  9 comps
CV           3.701    3.542    3.539    3.546    3.543    3.515    3.528    3.455    3.447    3.409
adjCV        3.701    3.542    3.539    3.546    3.543    3.515    3.528    3.454    3.447    3.408
       10 comps  11 comps  12 comps  13 comps  14 comps  15 comps  16 comps  17 comps  18 comps  19 comps
CV        3.409     3.398     3.382     3.392     3.393     3.386     3.381     3.398     3.405     3.408
adjCV     3.409     3.398     3.382     3.392     3.393     3.385     3.380     3.397     3.405     3.407
       20 comps  21 comps  22 comps  23 comps  24 comps  25 comps  26 comps  27 comps  28 comps  29 comps
CV        3.421      3.38      3.39     3.402     3.411     3.394     3.399     3.394     3.404     3.414
adjCV     3.421      3.38      3.39     3.401     3.412     3.394     3.399     3.394     3.404     3.414
       30 comps
CV         3.42
adjCV      3.42

TRAINING: % variance explained
   1 comps  2 comps  3 comps  4 comps  5 comps  6 comps  7 comps  8 comps  9 comps  10 comps  11 comps
X    10.31   18.593   24.538    29.66    34.52    39.17    43.45    47.51    51.37     54.99     58.41
Y     8.78    9.432    9.575    10.35    12.10    12.15    16.25    17.18    19.23     19.51     20.65
   12 comps  13 comps  14 comps  15 comps  16 comps  17 comps  18 comps  19 comps  20 comps  21 comps
X     61.76     64.95     68.04     70.89     73.60     76.26     78.81     81.21     83.56     85.79
Y     21.25     21.29     21.83     22.52     22.97     22.97     23.06     23.37     23.48     25.11
   22 comps  23 comps  24 comps  25 comps  26 comps  27 comps  28 comps  29 comps  30 comps
X     87.88     89.86     91.72     93.56     95.21     96.77     98.08     99.09    100.00
Y     25.16     25.16     25.19     26.28     26.52     27.11     27.17     27.17     27.29
```

Comment interpréter ces résultats ?

Les sorties dans la partie “validation” sont l’erreur quadratique moyenne de prédiction (RMSEP). Nous avons deux estimations de validation croisée : la CV qui est la validation croisée ordinaire et la adjCV qui est la validation croisée corrigée par un biais .

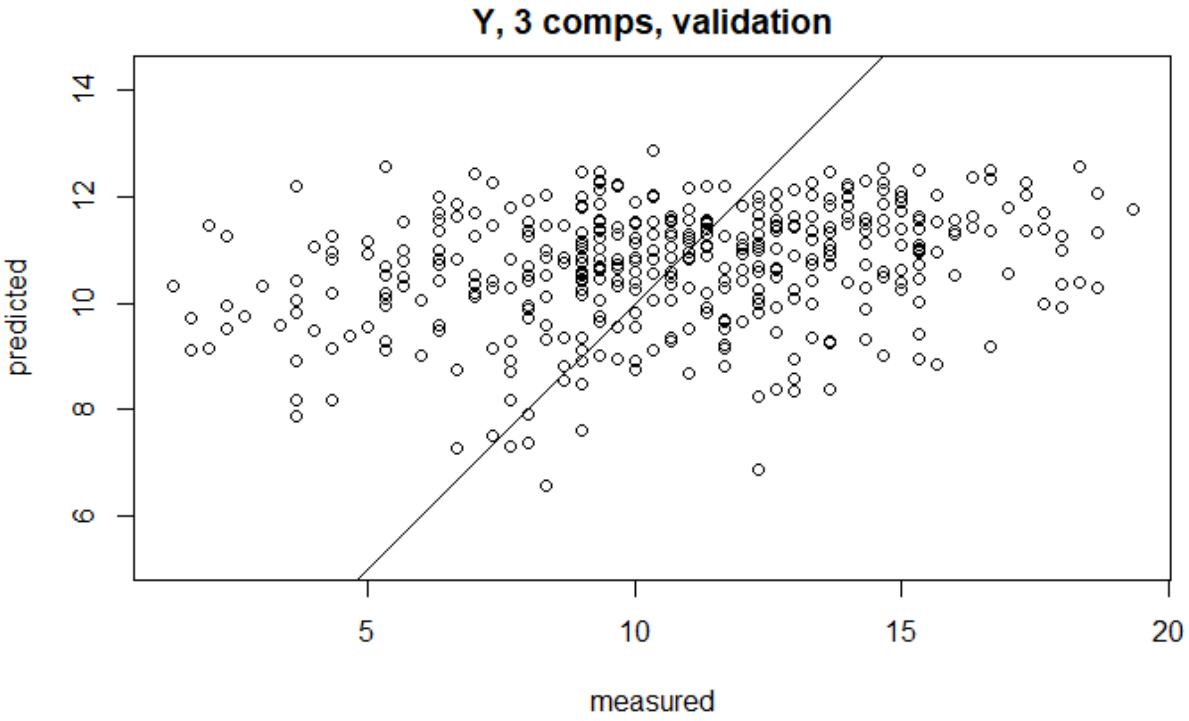
On remarque qu’il n’y a aucune différence entre les deux. Il semble que 3 composantes suffisent pour expliquer un peu plus de 24% de la variabilité des données. De plus, à partir de 3 composantes, le score CV diminue très faiblement.

Pour confirmer notre choix du nombre de composantes nous allons tracer le graphique de cette RMSEP.



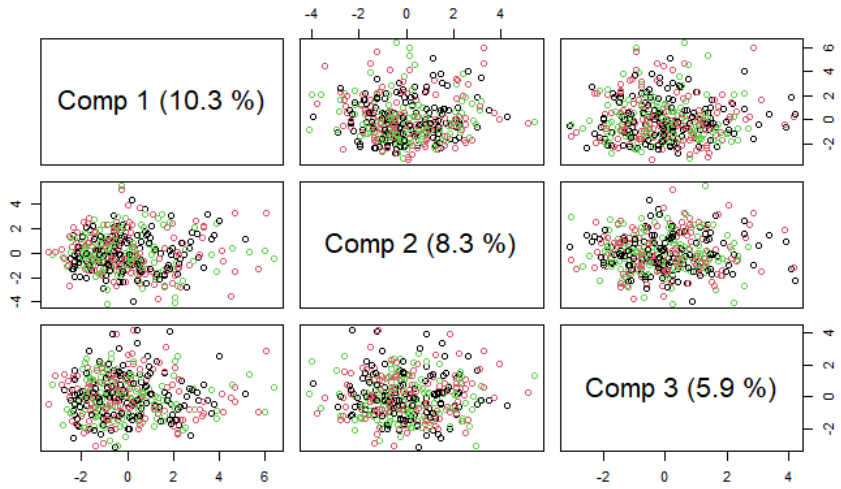
D’après la méthode du « coude », on remarque qu’effectivement 3 composantes suffisent pour expliquer nos observations.

A présent nous allons réaliser les prédictions croisées de la variable moyenne à partir de ces 3 composantes.



Cette prédiction croisée nous invite à penser que la moyenne de l'étudiant prédite sera principalement entre 8 et 15, soit des élèves moyen ou bon. Ceci semble cohérent avec notre jeu de données, car comme nous l'avons montré, une grande partie des élèves interrogés ont une moyenne entre 8 et 15.

Passons maintenant aux scores pour les trois composantes en effectuant un graphique de scores.



Cela nous donne un tracé par paires des valeurs de scores pour les 3 composantes. Ces graphiques de scores sont souvent utilisés pour rechercher des modèles, des groupes ou des valeurs aberrantes dans les données. Chaque diagramme des scores est un nuage de points de scores de X tracé à partir de deux composantes du modèle.

Par exemple, le premier diagramme sur la première colonne et la deuxième ligne représente le nuage de points des scores X tracé à partir des composantes 1 et 2. On remarque un certain regroupement des points pour chaque diagramme, ce qui peut se traduire par le fait que nos composantes expliquent moyennement notre modèle.

C'est sûrement dû à notre base de données.

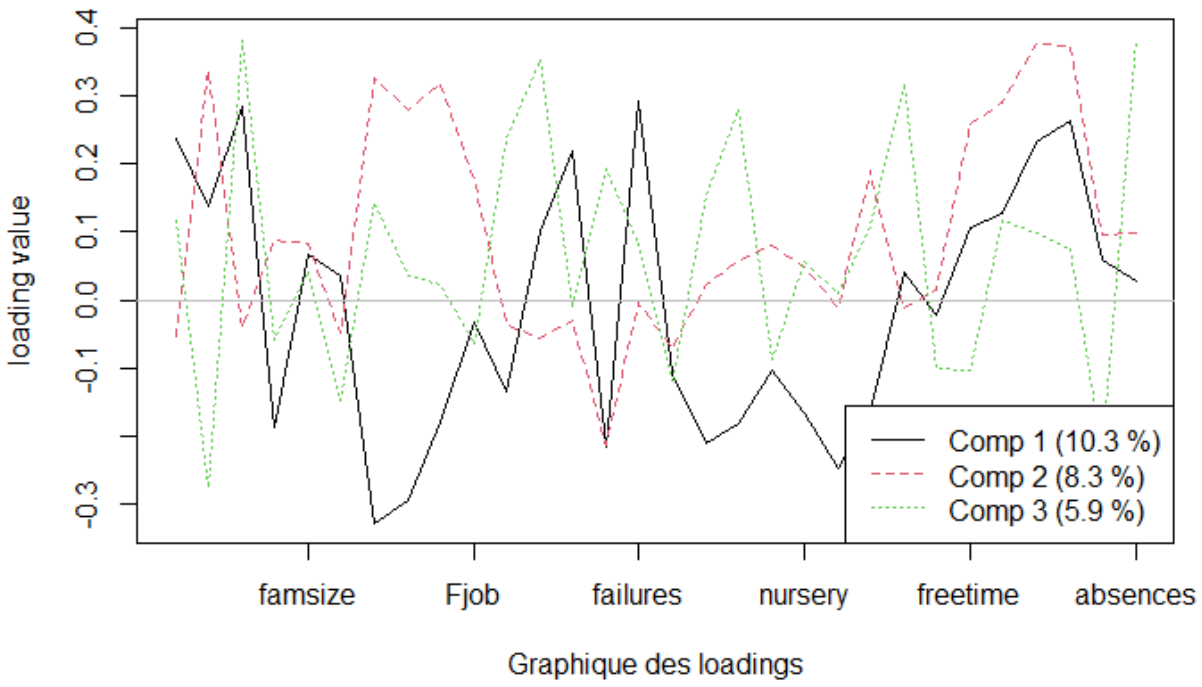
De plus, on remarque que certains points s'éloignent de ses regroupements. Cela représente des observations qui peuvent jouer un rôle significatif dans le modèle.

Pour la suite, nous allons analyser les corrélations entre les composantes principales et les variables explicatives. Cette relation de corrélation est reflétée par les loadings. Afin de faciliter l'analyse de ces loadings, nous allons les représenter graphiquement.

```
Loadings:
           Comp 1 Comp 2 Comp 3 Comp 4 Comp 5 Comp 6 Comp 7 Comp 8 Comp 9 Comp 10 Comp 11 Comp 12 Comp 13 Comp 14 Comp 15 Comp 16 Comp 17 Comp 18 Comp 19 Comp 20 Comp 21 Comp 22 Comp 23 Comp 24
school      0.237         0.117 -0.384  0.233 -0.133  0.193 -0.145                0.151          0.100  -0.144  -0.161   0.358   0.111                                0.108  -0.218  -0.132
sex         0.137  0.336 -0.277        0.135                       0.151         -0.259          -0.184                  -0.118                 -0.184  -0.273                 0.141  -0.222
age         0.284         0.381        0.198 -0.245                                                                    0.201                 0.183  -0.135         -0.100         -0.275
address    -0.186                0.471               0.145 -0.338                                                                    0.251         0.156                 -0.226  -0.547
famsize                                 0.114  0.288  0.472         0.250  0.232          0.159   0.254          0.135          0.161  0.314  0.391         -0.121  0.182  0.147   0.262
Pstatus                -0.149 -0.130 -0.215 -0.401 -0.192 -0.305  0.175  0.164                                 0.127   0.306 -0.375  0.252  0.100         -0.140  0.431
Medu       -0.327  0.324  0.143        0.233                                                -0.107                                                                0.214
Fedu       -0.295  0.277        -0.170  0.206        -0.120        -0.148         0.248  -0.144                 -0.183                 0.134  0.225  -0.168  0.175  0.139          0.110
Mjob       -0.178  0.319                0.268 -0.124                              -0.276   0.226          -0.148   0.263   0.241   0.104  -0.187         0.129                 -0.193  -0.137
Fjob               0.176        -0.172  0.256        -0.377 -0.169        -0.209  0.131   0.150   0.206   0.227   0.382  -0.215   0.165  0.237  -0.170  0.170  -0.231  -0.153          0.190
reason     -0.135         0.236        -0.103                0.131  0.471         -0.164          -0.379   0.174  -0.136          0.415  0.133         0.332  0.139  0.189  -0.179
guardian    0.101         0.353  0.219         -0.167        0.240 -0.147 -0.275  -0.217                 -0.260   0.165  -0.180  -0.233  0.110  -0.205  -0.111         0.341  -0.103   0.192
traveltime  0.219               -0.385                      0.201         0.167          0.345   0.226                 -0.372  -0.163  -0.132         0.225         0.182  -0.281  -0.306
studytime  -0.214 -0.213  0.192        -0.142 -0.201                                     0.116  -0.108           0.105   0.355          0.281  -0.458         -0.250  -0.278  0.124          -0.123
failures    0.292               0.257                -0.275  0.101                      0.185  -0.105  -0.196   0.268   0.117   0.175          0.109  0.210  0.406          0.376
schoolsup  -0.108        -0.120        -0.248  0.287 -0.208  0.299         -0.164  0.192   0.262   0.285                  0.493          0.119  -0.153  -0.269         0.108  -0.162
famsup     -0.209         0.154        -0.312        -0.194  0.105 -0.346  0.192  -0.102   0.238  -0.116  -0.180  -0.104  -0.169          0.178  0.333         -0.175  -0.103  0.236  -0.173
paid       -0.181         0.279 -0.199 -0.309         0.175 -0.172 -0.137         -0.244   0.163                  0.127                  0.268  -0.227  -0.181  0.153  -0.312  -0.220  0.166
activities -0.104                                -0.260 -0.110  0.364  0.484                -0.231   0.152  -0.245   0.175          -0.127  0.174  0.101  -0.155  0.134  -0.424          -0.116
nursery    -0.166                       0.104  0.177  0.209  0.182         0.356  0.295   0.244  -0.378   0.178   0.286  -0.353          -0.265  0.125                 0.117
higher     -0.247                0.235 -0.116         0.308  0.102         -0.222          -0.268   0.286  -0.105                 -0.122          -0.180  0.463  0.142          0.406  -0.144
internet   -0.162  0.189  0.109  0.181 -0.115 -0.115        -0.310  0.181  0.190           0.235   0.345          -0.208                 -0.407                 0.279                 0.250
romantic           0.318  0.108  0.135        -0.169                0.513  0.236  -0.223   0.254          -0.164          0.159  0.195  -0.410                 0.177
famrel                    0.132        -0.411  0.163  0.258 -0.110                0.227   0.176   0.623  -0.181                 0.151                 -0.101  0.194          0.167
freetime    0.106  0.257 -0.104  0.168 -0.138 -0.310  0.159  0.198         0.175   0.132          -0.253  -0.277          0.150          -0.206  0.191  -0.483  -0.126         0.164
goout      0.128  0.290  0.116  0.101 -0.292 -0.117  0.106                -0.140  0.476                          0.127  -0.137                 0.220  0.157  0.112          -0.271  0.136
Dalc       0.232  0.374        -0.151 -0.231  0.147                                                             0.107   0.152          -0.176  -0.160                 0.192  -0.223
Walc       0.262  0.372        -0.130 -0.286  0.168                                -0.191          0.131
health             -0.225                              0.233 -0.395  0.393  -0.270  -0.362          0.180   0.171   0.189   0.109                 0.177                 -0.308   0.137
absences           0.384  0.151         0.201 -0.220  0.167                -0.200  -0.104   0.130   0.291  -0.192   0.180  -0.380          0.133  0.143  -0.343  -0.209
```

```
               Comp 1 Comp 2 Comp 3 Comp 4 Comp 5 Comp 6 Comp 7 Comp 8 Comp 9 Comp 10 Comp 11 Comp 12 Comp 13 Comp 14 Comp 15 Comp 16 Comp 17 Comp 18 Comp 19 Comp 20 Comp 21 Comp 22 Comp 23 Comp 24
SS loadings     1.000  1.000  1.000  1.000  1.000  1.000  1.000  1.000  1.000  1.000  1.000  1.000  1.000  1.000  1.000  1.000  1.000  1.000  1.000  1.000  1.000  1.000  1.000  1.000
Proportion Var  0.033  0.033  0.033  0.033  0.033  0.033  0.033  0.033  0.033  0.033  0.033  0.033  0.033  0.033  0.033  0.033  0.033  0.033  0.033  0.033  0.033  0.033  0.033  0.033
Cumulative Var  0.033  0.067  0.100  0.133  0.167  0.200  0.233  0.267  0.300  0.333  0.367  0.400  0.433  0.467  0.500  0.533  0.567  0.600  0.633  0.667  0.700  0.733  0.767  0.800
               Comp 25 Comp 26 Comp 27 Comp 28 Comp 29 Comp 30
SS loadings      1.000   1.000   1.000   1.000   1.000   1.000
Proportion Var   0.033   0.033   0.033   0.033   0.033   0.033
Cumulative Var   0.833   0.867   0.900   0.933   0.967   1.000
```

	Comp 25	Comp 26	Comp 27	Comp 28	Comp 29	Comp 30
school	-0.195			-0.531		
sex	-0.218	0.467	0.382			
age	-0.208			0.600	-0.111	-0.135
address	-0.187	-0.192	0.130	-0.128	0.117	
famsize						
Pstatus						
Medu	0.153	-0.130	0.158	-0.141	-0.514	-0.483
Fedu	-0.110	-0.223	0.169	0.186	0.426	0.351
Mjob	0.389	0.172	-0.301		0.260	0.138
Fjob	-0.228					
reason	-0.182					
guardian	-0.306			-0.189		
traveltime		-0.208	0.166	0.166		
studytime			0.301	-0.131	0.157	
failures	0.206	-0.140	0.331			
schoolsup	-0.107	0.170		0.123		-0.107
famsup	-0.194	0.261	-0.246	-0.135		
paid	0.241		0.336	0.124		
activities		-0.146	-0.148			
nursery	-0.170		-0.162			
higher		0.160		0.148		
internet	-0.347					
romantic	0.207	0.208				
famrel	0.169		-0.103	-0.154		
freetime		-0.221	0.167	0.175		
goout		0.423			-0.157	0.246
Dalc		-0.350	-0.239		-0.349	0.415
Walc			-0.132		0.482	-0.565
health	-0.257					
absences			0.259	-0.183		



On considère qu'une valeur supérieure à 0,4 en valeur absolue, ou presque est l'indication d'une liaison significative.

La première composante (axe 1) est saturée de façon plus importante pour les variables Medu (-0.327) et Fedu (- 0,295).

Pour la saturation de la composante 2, on retrouve Dalc (0,374) et Walc (0,372).

Enfin, la composante 3 sature de façon plus importante les variables age (0,381), guardian (0,353), romantic(0,318) et absences (0,384). Graphiquement, plus la courbe est proche de 0, moins la variable correspondante est saturée.

Passons maintenant à la corrélation entre la variable à expliquer qualité et chaque composante.

```
Loadings:
  Comp 1 Comp 2 Comp 3 Comp 4 Comp 5 Comp 6 Comp 7 Comp 8 Comp 9 Comp 10 Comp 11 Comp 12 Comp 13 Comp 14
Y -0.623  0.189 -0.105 -0.263  0.404         0.659 -0.324  0.491 -0.188  -0.391  -0.286           0.282
  Comp 15 Comp 16 Comp 17 Comp 18 Comp 19 Comp 20 Comp 21 Comp 22 Comp 23 Comp 24 Comp 25 Comp 26
Y -0.333  -0.275          -0.125  -0.241  -0.149  -0.578                          -0.521  -0.261
  Comp 27 Comp 28 Comp 29 Comp 30
Y  0.414  -0.142          -0.248

               Comp 1 Comp 2 Comp 3 Comp 4 Comp 5 Comp 6 Comp 7 Comp 8 Comp 9 Comp 10 Comp 11 Comp 12
SS loadings     0.388  0.036  0.011  0.069  0.164  0.005  0.435  0.105  0.241   0.035   0.153   0.082
Proportion Var  0.388  0.036  0.011  0.069  0.164  0.005  0.435  0.105  0.241   0.035   0.153   0.082
Cumulative Var  0.388  0.424  0.435  0.504  0.667  0.673  1.108  1.213  1.454   1.489   1.642   1.723
               Comp 13 Comp 14 Comp 15 Comp 16 Comp 17 Comp 18 Comp 19 Comp 20 Comp 21 Comp 22 Comp 23
SS loadings      0.005   0.079   0.111   0.075   0.001   0.016   0.058   0.022   0.334   0.009   0.002
Proportion Var   0.005   0.079   0.111   0.075   0.001   0.016   0.058   0.022   0.334   0.009   0.002
Cumulative Var   1.729   1.808   1.919   1.995   1.995   2.011   2.069   2.091   2.425   2.434   2.436
               Comp 24 Comp 25 Comp 26 Comp 27 Comp 28 Comp 29 Comp 30
SS loadings      0.006   0.271   0.068   0.171   0.020   0.002   0.061
Proportion Var   0.006   0.271   0.068   0.171   0.020   0.002   0.061
Cumulative Var   2.441   2.712   2.780   2.952   2.972   2.974   3.036
```

La variable moyenne est fortement corrélé à la composante 1 (-0,623) tandis qu'elle est faiblement corrélée avec les composantes 2 et 3 (respectivement 0,189 et -0,105).

On peut donc faire une estimation des coefficients de régression de nos variables pour la variable moyenne.

Ces coefficients nous informent de l'impact de chaque variable explicative sur la moyenne. Par exemple, si nous rajoutons une unité d'absence, la moyenne sera améliorée de 15,8% environ.

De même, rajouter une unité de la variable traveltime diminuerait la moyenne de 16% environ.

Les quatres variables qui impactent le plus la moyenne sont failures (-0,95), goout (-0,57), schoolsup (-0,51) et sex (0,48).

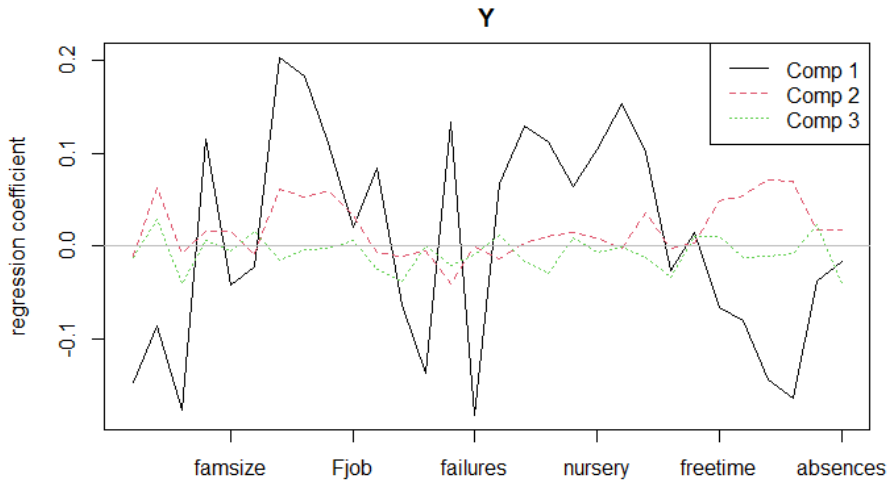
Cependant, même si l'estimation de ces coefficients nous apportent une solution pour améliorer la moyenne d'un étudiant, celle-ci comporte quelques limites.

En effet, rajouter des unités des variable favorable à l'amélioration augmentera la moyenne mais plus cette méthode sera utilisée et moins elle sera efficace.

Comme comparaison on peut prendre le temps de travail d'un étudiant. Travailler ses cours est tout à fait sérieux et est bénéfique mais si on travaille durant plusieurs dizaines d'heures de suite la courbe de progression et de compréhension vont croître de moins en moins jusqu'à stagner.

```
, , 30 comps

                Y
school      0.085522381
sex         0.481893502
age        -0.186492381
address     0.227256123
famsize     0.287106061
Pstatus    -0.090642146
Medu        0.378675125
Fedu        0.165690969
Mjob       -0.227299202
Fjob        0.203586300
reason      0.295570152
guardian    0.035735554
traveltime -0.160148763
studytime   0.461802337
failures   -0.957482853
schoolsup  -0.511744764
famsup     -0.388329398
paid        0.065976129
activities -0.075304463
nursery    -0.007805372
higher      0.286936264
internet    0.184950246
romantic   -0.324667911
famrel     -0.007415180
freetime    0.280961709
goout      -0.571754500
Dalc       -0.089335068
Walc        0.031462981
health     -0.161614006
absences    0.158607639
```

Pour faciliter la compréhension de cette régression nous allons encore une fois représenter graphiquement la régression effectuée sur chaque variable pour prédire la variable moyenne, et ce pour chacune des trois composantes retenues.



Graphique des coefficients de régression

Nous avons en abscisse chacune de nos variables explicatives.

On observe, par exemple, pour la variable failures qu'elle impact négativement la moyenne d'un élève dans la composante 1 et neutre pour les composantes 2 et 3.

Cette PCR nous a permis de déterminer les variables qui impactent le plus la moyenne d'un étudiant à partir d'une régression linéaire sur nos composantes principales. Cependant, les résultats obtenus ne semblent pas satisfaisants.

Notre base de données présente une multicolinéarité entre certaines variables. Par exemple, le fait qu'un élève fasse ou non une ou plusieurs activités extrascolaires dépend naturellement de la variable freetime (du temps libre dont cet étudiant dispose après les cours).

Afin d'estomper cette multicolinéarité qui fausse notre étude, nous allons appliquer une régression par la méthode des moindres carrés partiels (PLS).

c. Régression des moindres carrés partiels

Le principal avantage de cette régression (ou PLS) et ce qui lui donne une forte supériorité face à la PCR est la possibilité d'avoir des variables explicatives fortement corrélées, ce qui est le cas dans notre base de données.

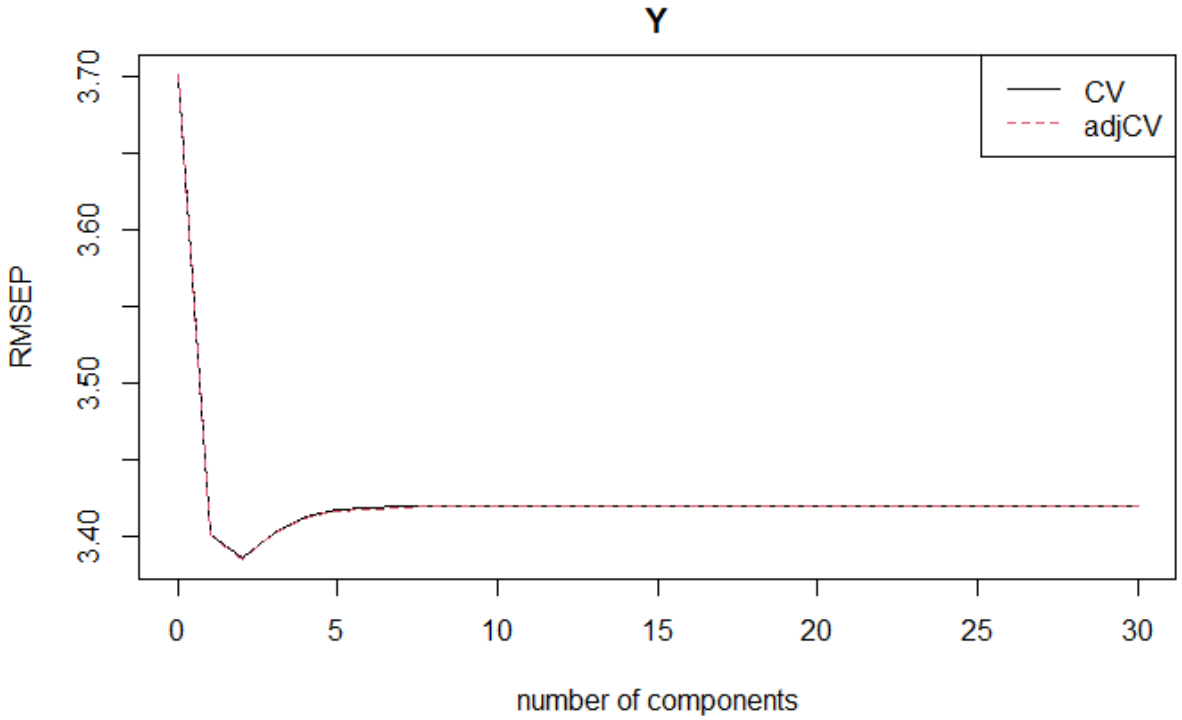
Commençons donc par rechercher ces composantes principales nécessaires pour la régression.

```
Data:   X dimension: 395 30
        Y dimension: 395 1
Fit method: kernelpls
Number of components considered: 30

VALIDATION: RMSEP
Cross-validated using 395 leave-one-out segments.
       (Intercept)  1 comps  2 comps  3 comps  4 comps  5 comps  6 comps  7 comps  8 comps  9 comps
CV           3.701    3.402    3.386    3.403    3.413    3.417    3.419    3.420     3.42     3.42
adjCV        3.701    3.402    3.386    3.403    3.412    3.417    3.418    3.419     3.42     3.42
       10 comps  11 comps  12 comps  13 comps  14 comps  15 comps  16 comps  17 comps  18 comps
CV         3.42      3.42      3.42      3.42      3.42      3.42      3.42      3.42      3.42
adjCV      3.42      3.42      3.42      3.42      3.42      3.42      3.42      3.42      3.42
       19 comps  20 comps  21 comps  22 comps  23 comps  24 comps  25 comps  26 comps  27 comps
CV         3.42      3.42      3.42      3.42      3.42      3.42      3.42      3.42      3.42
adjCV      3.42      3.42      3.42      3.42      3.42      3.42      3.42      3.42      3.42
       28 comps  29 comps  30 comps
CV         3.42      3.42      3.42
adjCV      3.42      3.42      3.42

TRAINING: % variance explained
   1 comps  2 comps  3 comps  4 comps  5 comps  6 comps  7 comps  8 comps  9 comps  10 comps  11 comps
X    8.995    14.19    18.88    24.78    28.73    32.20    35.78    39.17    42.18     45.49     48.85
Y   20.723    26.42    27.09    27.25    27.28    27.29    27.29    27.29    27.29     27.29     27.29
   12 comps  13 comps  14 comps  15 comps  16 comps  17 comps  18 comps  19 comps  20 comps  21 comps
X     51.90     54.83     57.10     60.22     63.09     65.69     68.86     71.71     74.21     77.06
Y     27.29     27.29     27.29     27.29     27.29     27.29     27.29     27.29     27.29     27.29
   22 comps  23 comps  24 comps  25 comps  26 comps  27 comps  28 comps  29 comps  30 comps
X     79.80     82.38     85.27     87.68     90.14     92.94     95.09     97.34    100.00
Y     27.29     27.29     27.29     27.29     27.29     27.29     27.29     27.29     27.29
```

De la même façon que lors de la PCR, les résultats de validation de la RMSEP nous informe sur le nombre de composantes principales nécessaires. On remarque qu'il n'y a aucune différence entre la CV et l'adjCV. Il semble qu'uniquement deux composantes expliquent un peu plus de 14% de la variabilité des données.

De plus, à partir de 7 composantes, le score CV reste constant (à 3,42), ce qui pourrait nous inviter à ajouter de nouvelle composantes.

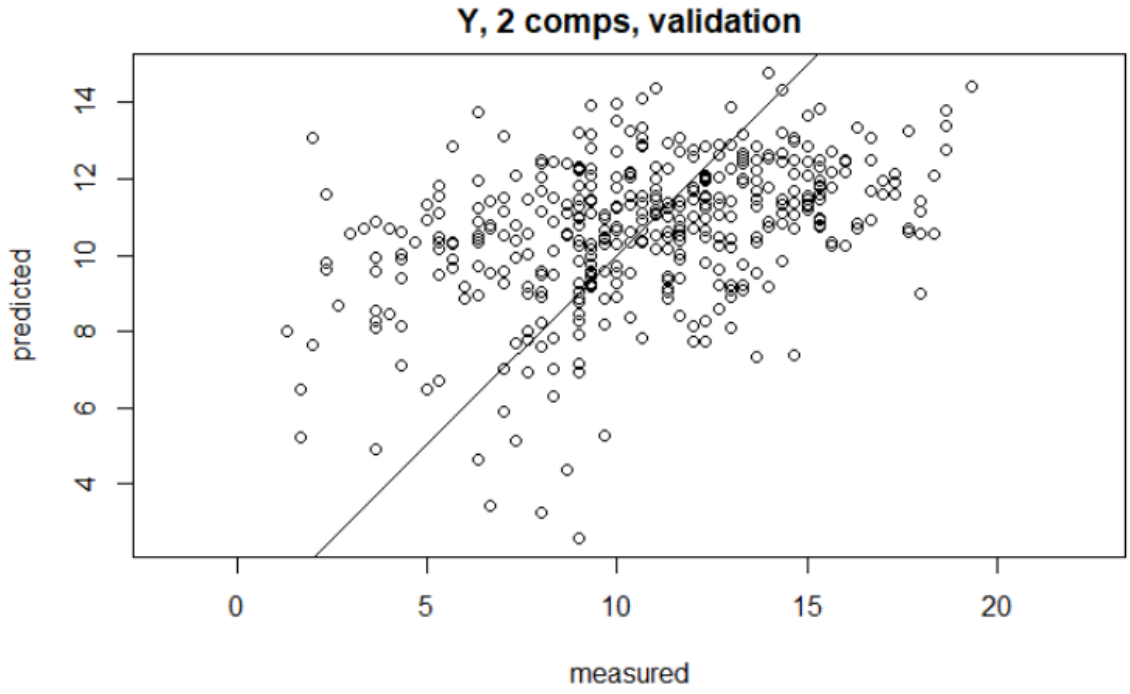
Effectuons un graphique de cette RMSEP, afin de confirmer notre choix du nombre de composantes.



Ce graphique nous invite à choisir au maximum deux composantes principales. En effet, on remarque en appliquant la méthode du « coude » qu'à partir de deux composantes, en rajouter une nouvelle n'apporterait pas de changement significatif.

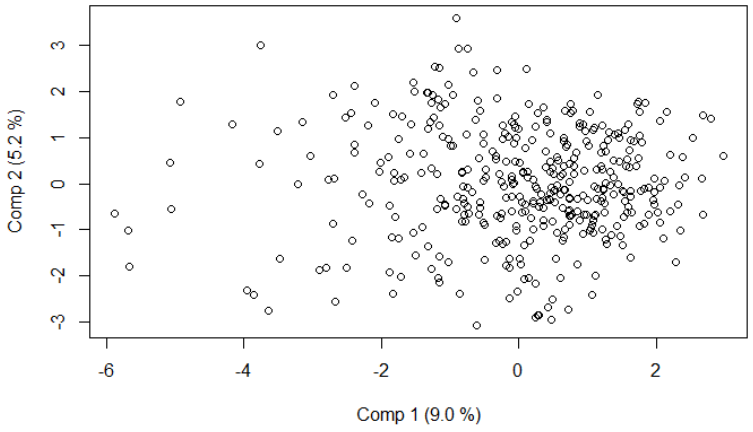
Nous allons donc poursuivre notre étude avec deux composantes.

Pour la suite, nous allons effectuer des prédictions croisées de la variable qualité à partir de ces deux composantes.



Cette prédiction croisée nous invite à penser, comme dans notre PCR, que la moyenne d'un étudiant prédite sera principalement entre 9 et 16, soit des élèves moyen et bon. On remarque cependant que, contrairement à la PCR, certaines moyennes de 4 ou 5 se rapprochent de la prédiction (plus proche de la gauche). La PLS semble prédire certaines moyennes moins bonnes que la PCR.

Analysons maintenant les scores pour les deux composantes en effectuant un graphique de scores.



Ce diagramme représente le nuage de points des scores X tracé à partir de la composante 1 et de la composante 2. On remarque qu'il n'y a pas de regroupement significatif des points, ce qui se traduit par le fait que la PLS semble plus adaptée à nos observations.

Voyons maintenant les corrélations entre les composantes principales et les variables explicatives à partir des loadings.

Loadings:

	Comp 1	Comp 2	Comp 3	Comp 4	Comp 5	Comp 6	Comp 7	Comp 8	Comp 9	Comp 10	Comp 11
school	-0.166	0.321	-0.281	-0.108	0.442	0.195	-0.243	0.159	-0.176		-0.235
sex		0.338	-0.276	0.375		-0.172	0.218	-0.147		-0.156	
age	-0.275	0.238		-0.109		0.301	-0.459	0.259		0.127	0.383
address	0.192	-0.130	0.101		-0.151	0.392		-0.419	0.260	-0.272	0.141
famsize		0.256	-0.201			0.373	-0.182	-0.244	0.176		
Pstatus							0.205	-0.285		0.372	-0.109
Medu	0.405	-0.183	-0.244	0.283		0.203		0.378			
Fedu	0.351	-0.205	-0.255	0.265	0.131		-0.197		-0.328	-0.127	
Mjob	0.255	-0.115	-0.538		-0.150		-0.157	0.104			0.196
Fjob			-0.145	0.280	0.235	-0.218		0.103	-0.170	0.278	0.133
reason	0.158		0.205		-0.269	0.180		0.107	-0.422	0.237	
guardian	-0.144		0.232			0.313	-0.401	0.146			0.315
traveltime	-0.231	0.119	-0.112		0.293	-0.205		0.189	-0.158		-0.365
studytime	0.186		0.433	-0.123	0.167	0.160	-0.232			0.128	-0.222
failures	-0.453	-0.203	-0.133			0.335			-0.211	-0.194	0.111
schoolsup		-0.427	0.120		0.148		0.138	-0.212		0.258	
famsup		-0.466	0.155	0.127			-0.148	0.190		0.148	-0.138
paid	0.184	-0.131			-0.168	0.318	-0.152		-0.184		-0.428
activities	0.116		-0.149				0.193	-0.120	-0.172	0.223	
nursery	0.167	-0.156	-0.122		0.212	0.126	-0.300		0.229		
higher	0.305			-0.138					-0.104	0.105	
internet	0.184	-0.112		0.173	-0.266	0.271		-0.126			
romantic	-0.104	-0.112			-0.118		-0.377	0.398	0.104	-0.121	
famrel							0.132	0.111	0.154	-0.515	
freetime		0.103		0.421			-0.125	-0.198	0.451	-0.108	
goout	-0.154		-0.301	0.228	-0.348	0.230	-0.162	-0.159	0.125		-0.250
Dalc	-0.168	0.184	-0.248	0.407	-0.350		-0.200	-0.108	-0.166	0.219	-0.195
Walc	-0.203	0.196	-0.254	0.404	-0.319	0.257	0.273	0.194		0.288	-0.195
health				0.214	0.143	-0.209	0.228		-0.155	-0.242	0.418
absences			0.245	0.186	-0.330	0.178	-0.193	0.332	-0.363		0.120

	Comp 1	Comp 2	Comp 3	Comp 4	Comp 5	Comp 6	Comp 7	Comp 8	Comp 9	Comp 10	Comp 11	Comp 12
SS loadings	1.171	1.087	1.332	1.149	1.186	1.276	1.293	1.198	1.109	1.127	1.148	1.110
Proportion Var	0.039	0.036	0.044	0.038	0.040	0.043	0.043	0.040	0.037	0.038	0.038	0.037
Cumulative Var	0.039	0.075	0.120	0.158	0.198	0.240	0.283	0.323	0.360	0.398	0.436	0.473

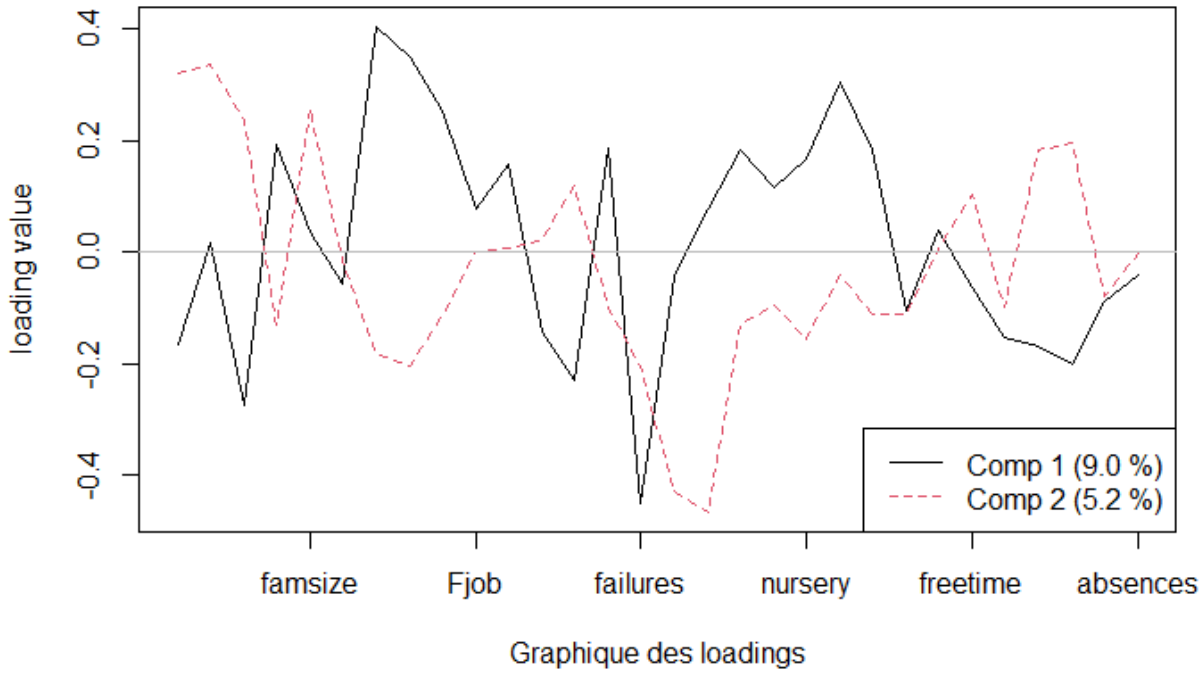
	Comp 13	Comp 14	Comp 15	Comp 16	Comp 17	Comp 18	Comp 19	Comp 20	Comp 21	Comp 22	Comp 23
SS loadings	1.115	1.122	1.160	1.241	1.110	1.058	1.058	1.075	1.024	1.018	1.036
Proportion Var	0.037	0.037	0.039	0.041	0.037	0.035	0.035	0.036	0.034	0.034	0.035
Cumulative Var	0.510	0.547	0.586	0.627	0.664	0.700	0.735	0.771	0.805	0.839	0.873

	Comp 24	Comp 25	Comp 26	Comp 27	Comp 28	Comp 29	Comp 30
SS loadings	1.012	1.021	1.031	1.009	1.052	6.211	7772.202
Proportion Var	0.034	0.034	0.034	0.034	0.035	0.207	259.073
Cumulative Var	0.907	0.941	0.976	1.009	1.044	1.251	260.325

Nous avons choisi d'afficher ici seulement les 11 premières composantes puisqu'on s'intéresse au deux premières.

Tout comme pour la PCR, nous allons considérer qu'une valeur supérieure à 0,35 (en valeur absolue) est l'indication d'une liaison significative. On remarque donc que la composante 1 sature de façon plus importante les variables Medu (0,405) et failures (-0,453). Tandis que pour la composante 2, on retrouve famsup (-0,466) et schoolsup (-0,427).

Dans le but d'avoir une vue d'ensemble de ces loadings, représentons graphiquement pour chaque composante.



Graphique des loadings

Graphiquement, plus la courbe est proche de 0, moins la variable correspondante est saturée.

On peut voir que la taille de la famille (variable famsize) est corrélée positivement avec la composante 2 mais n'est pas du tout corrélée avec la première.

Étudions alors le poids de chaque variable sur nos deux composantes.

```
Loadings:
           Comp 1 Comp 2
school              0.251
sex         0.146  0.313
age        -0.195  0.194
address     0.155
famsize     0.119  0.199
Pstatus
Medu        0.325 -0.196
Fedu        0.254 -0.233
Mjob        0.144 -0.268
Fjob
reason      0.174
guardian   -0.103
traveltime -0.186  0.109
studytime   0.195
failures   -0.544 -0.220
schoolsup  -0.199 -0.374
famsup             -0.404
paid        0.130 -0.133
activities         -0.137
nursery            -0.176
higher      0.274
internet    0.149
romantic   -0.149 -0.108
famrel
freetime            0.165
goout      -0.224 -0.168
Dalc       -0.105  0.152
Walc       -0.127  0.182
health     -0.116
absences
```

	Comp 1	Comp 2	Comp 3	Comp 4	Comp 5	Comp 6	Comp 7	Comp 8	Comp 9	Comp 10	Comp 11	Comp 12
SS loadings	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
Proportion Var	0.033	0.033	0.033	0.033	0.033	0.033	0.033	0.033	0.033	0.033	0.033	0.033
Cumulative Var	0.033	0.067	0.100	0.133	0.167	0.200	0.233	0.267	0.300	0.333	0.367	0.400

	Comp 13	Comp 14	Comp 15	Comp 16	Comp 17	Comp 18	Comp 19	Comp 20	Comp 21	Comp 22	Comp 23
SS loadings	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
Proportion Var	0.033	0.033	0.033	0.033	0.033	0.033	0.033	0.033	0.033	0.033	0.033
Cumulative Var	0.433	0.467	0.500	0.533	0.567	0.600	0.633	0.667	0.700	0.733	0.767

Ici, chaque coefficient nous indique pour chaque variable son poids dans la composante correspondante. On tient à souligner que les variables Pstatus, Fjob, famsup, activities, famrel, freetime et absences n'interviennent pas dans la composante 1.

De même, beaucoup de variables sont absentes dans la composante 2 comme le statut de cohabitation des parents d'un étudiant, le type d'adresse de l'étudiant, l'emploi du père, les raisons du choix de cette école par l'élève ou encore le temps passé à étudier durant la semaine.

Passons maintenant à l'observation des corrélations entre la variable à expliquer moyenne et chaque composante.

```
Loadings:
  Comp 1 Comp 2 Comp 3 Comp 4 Comp 5 Comp 6 Comp 7 Comp 8 Comp 9 Comp 10 Comp 11 Comp 12 Comp 13
Y 1.109  0.737  0.295  0.117
  Comp 14 Comp 15 Comp 16 Comp 17 Comp 18 Comp 19 Comp 20 Comp 21 Comp 22 Comp 23 Comp 24 Comp 25
Y
  Comp 26 Comp 27 Comp 28 Comp 29 Comp 30
Y

               Comp 1 Comp 2 Comp 3 Comp 4 Comp 5 Comp 6 Comp 7 Comp 8 Comp 9 Comp 10 Comp 11 Comp 12
SS loadings     1.229  0.544  0.087  0.014  0.005  0.001   0.00   0.00   0.00    0.00    0.00    0.00
Proportion Var  1.229  0.544  0.087  0.014  0.005  0.001   0.00   0.00   0.00    0.00    0.00    0.00
Cumulative Var  1.229  1.772  1.860  1.873  1.878  1.879   1.88   1.88   1.88    1.88    1.88    1.88
               Comp 13 Comp 14 Comp 15 Comp 16 Comp 17 Comp 18 Comp 19 Comp 20 Comp 21 Comp 22 Comp 23
SS loadings       0.00    0.00    0.00    0.00    0.00    0.00    0.00    0.00    0.00    0.00    0.00
Proportion Var    0.00    0.00    0.00    0.00    0.00    0.00    0.00    0.00    0.00    0.00    0.00
Cumulative Var    1.88    1.88    1.88    1.88    1.88    1.88    1.88    1.88    1.88    1.88    1.88
               Comp 24 Comp 25 Comp 26 Comp 27 Comp 28 Comp 29 Comp 30
SS loadings       0.00    0.00    0.00    0.00    0.00    0.00    0.00
Proportion Var    0.00    0.00    0.00    0.00    0.00    0.00    0.00
Cumulative Var    1.88    1.88    1.88    1.88    1.88    1.88    1.88
```

On remarque aisément que la variable moyenne a uniquement un lien avec les quatres premières composantes. Et est particulièrement corrélé à la première.

Effectuons maintenant une estimation des coefficients de régression de nos variables pour la variable moyenne.

```
, , 30 comps

                   Y
school      0.085522381
sex         0.481893502
age        -0.186492381
address     0.227256123
famsize     0.287106061
Pstatus    -0.090642146
Medu        0.378675125
Fedu        0.165690969
Mjob       -0.227299202
Fjob        0.203586300
reason      0.295570152
guardian    0.035735554
traveltime -0.160148763
studytime   0.461802337
failures   -0.957482853
schoolsup  -0.511744764
famsup     -0.388329398
paid        0.065976129
activities -0.075304463
nursery    -0.007805372
higher      0.286936264
internet    0.184950246
romantic   -0.324667911
famrel     -0.007415180
freetime    0.280961709
goout      -0.571754500
Dalc       -0.089335068
Walc        0.031462981
health     -0.161614006
absences    0.158607639
```

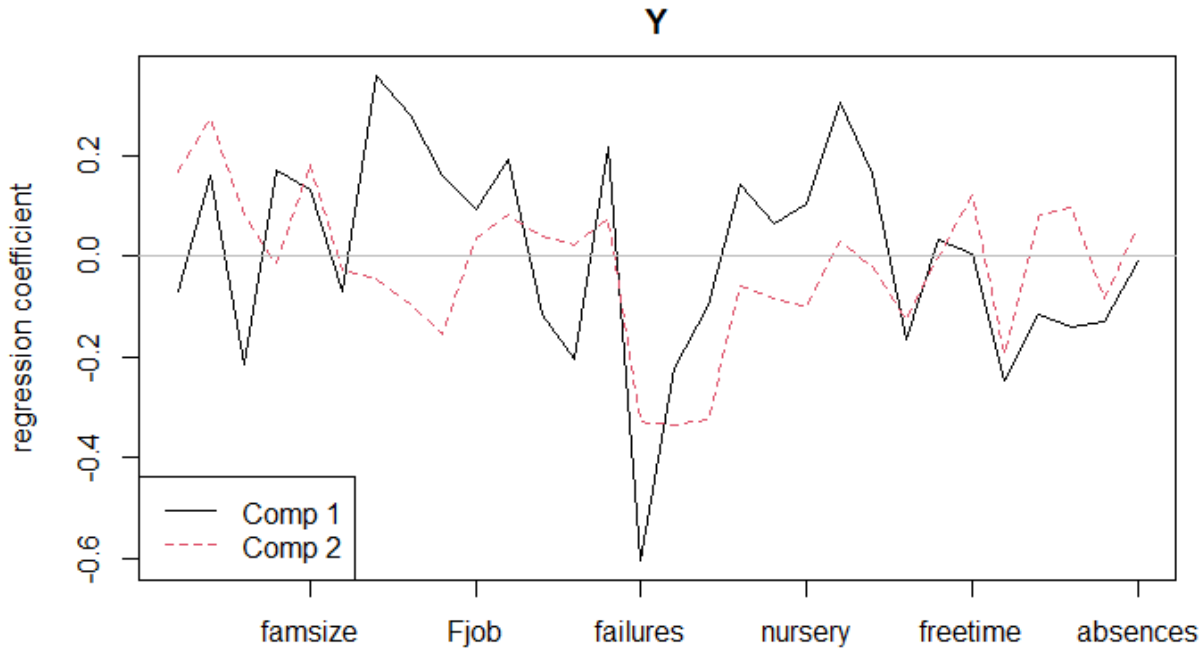
Ces coefficients nous informent de l'impact de chaque variable explicative sur la variable moyenne.

Ils s'interprètent de la même manière que dans la PCR.

Par exemple, si nous rajoutons une unité de temps libre, la moyenne de l'élève sera améliorée de 28% environ.

On remarque que les variables qui impactent le plus la moyenne d'un élève sont les mêmes que celles retenues pour la PCR.

Enfin pour faciliter la compréhension de cette régression nous allons représenter graphiquement la régression effectuée sur chaque variable pour prédire la variable moyenne, et ce pour chacune des trois composantes retenues.



Graphique des coefficients de regression

Comme précédemment nous avons en abscisse chacune de nos variables explicatives .

On observe, par exemple, pour la variable Fjob qu'elle impacte positivement sur la moyenne d'un étudiant dans la composante 1 et n'impacte pas la composante 2.

5. Conclusion de l’étude

L’objet de cette étude était de prédire la recette pour qu’un étudiant ait une bonne moyenne, à partir de 365 observations. Nous avons commencé par effectuer une régression linéaire multiple sur les variables explicatives. Nous avons constaté que notre jeu de données suivait une distribution normale, et que cette régression était en adéquation avec nos observations cependant il y avait peut-être une méthode plus adaptée pour répondre à notre problématique.

Par la suite nous avons donc choisi d’appliquer une Régression sur les Composantes Principales (PCR). Les composantes principales étant des combinaisons linéaires de nos variables. Procéder de la sorte nous permettait de réduire la dimension des variables à analyser. Nous avons estimé que trois composantes suffisaient à donner une bonne représentation de nos variables. À partir de cette régression, via les composantes principales, nous avons pu distinguer quelles variables explicatives impactaient le plus la moyenne d’un étudiant. Il en ressorti que si un élève avait un nombre d'échecs scolaires antérieurs élevé cela diminuait considérablement sa moyenne.Cependant, même si cette estimation nous apportait une solution pour accroître la moyenne, la multicolinéarité présente dans notre base de données faussait les résultats en plus du fait que certaines corrélations soient assez surprenantes. En effet, la variable schoolsup influencait négativement le modèle alors que l’on pourrait penser qu’un soutien scolaire supplémentaire aide un élève à améliorer sa moyenne. Augmenter la quantité d’alcool influerait donc sur la qualité mais également sur la densité.

Afin d’estomper cette multicolinéarité et rendre notre estimation plus pertinente, nous avons appliqué une régression par la méthode des Moindres Carrés Partiels (PLS). Le principal avantage de cette méthode est la possibilité d’avoir des variables explicatives fortement corrélées. En effet, les composantes principales choisies sont orthogonales les unes aux autres. À partir de cette régression, nous avons découvert qu’augmenter le temps de travail d’un étudiant améliorait la moyenne de celui-ci.

Même si nous avons trouvé des leviers qui permettent d'améliorer la moyenne d'un élève dans l'enseignement secondaire, il est important de souligner que la conclusion de notre étude comporte certaines limites.

Tout d'abord, une des limites la plus flagrante est que nous avons conclu qu'augmenter le temps de travail d'un étudiant augmenterait sa moyenne. Il paraît évident que la quantité de travail n'est pas la seule chose à prendre en compte et que plusieurs autres aspects comme le degré de concentration de l'élève ou ses facilités jouent aussi un rôle important.

De ce fait, l'absence de ces informations dans notre base de données rend notre analyse incomplète.

Nous terminons ici notre étude, en citant une femme qui à marqué à jamais les sciences et plus précisément la physique: "La vie n'est facile pour aucun de nous. Mais quoi, il faut avoir de la persévérance, et surtout de la confiance en soi. Il faut croire que l'on est doué pour quelque chose, et que, cette chose, il faut l'atteindre coûte que coûte."