

Geoducation

Antoine Drouhin, Aurélien Garret, Cécile Hu, Lucas Morel

TODO, Note/CR de Réunion avec la prof :

- décrire la base de données : taille, variable quantitatif ou qualitatif (à faire pour la présentation orale)
- Ségrégation scolaire
- Idée :
- Essayer de faire une classification sur les différents taux de réussite et les filières (pas étonné qu'en ts gros tx de réussite, qu'en L non etc,...) Essayer de voir pourquoi meilleur on est meilleur on sera et vice versa.
- Choisir un indicateur de réussite, créer un indicateur en additionnant les taux de réussite par lycée, puis régression pouvoir si dépend de la filière et de la géographie.
- Secteur privé/public
- Rural/urbain
- Puis régression synthétique
- Cherche taille des communes pour joindre

Introduction

Base de données

Notre de base de données a été trouvée sur le site Data.gouv. Nous avons croisé deux jeux de données distincts. Le premier concerne des données sur la performance des lycées en France (taux de réussite etc). Le second présente des données géographiques pour l'ensemble des établissements scolaires français (Coordonnées GPS, etc.).

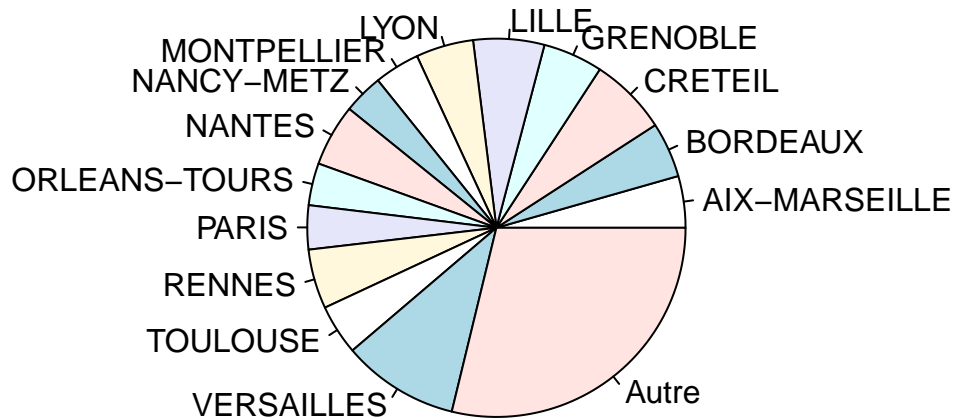
<https://www.data.gouv.fr/fr/>

Nous avons choisi cette base car elle présente une approche intéressante sur la compréhension d'un sujet qui nous concerne tous, l'éducation. L'approche géographique des questions de réussite scolaire nous a semblé être un champ suffisamment complet pour permettre l'utilisation des méthodes d'analyse vue en cours.

La base de données comporte de nombreuses données qui sont réparties sur toutes la France. Nous avons des répartitions de données par établissements, villes, départements et académies. Par ailleurs l'ensemble des bacs généraux et technologiques ainsi que certains bacs professionnels sont représentés.

```
print(pie(slices, labels = lbls, main="Répartition des effectifs par académie"))
```

Répartition des effectifs par académie



NULL

L'intérêt et le questionnement général porte sur la réussite scolaire de la France en fonction de la position géographique des établissements scolaires. Y a-t-il une corrélation entre la situation géographique des établissements et la réussite scolaire des étudiants ?

Plusieurs problématiques en découlent comme :

Y a-t-il des différences de réussites entre le top 10 des grandes villes en France et les villes de province ? Paris bénéficie-t-elle d'une réussite supérieure au reste de la France ? Quelles régions de France semblent réussir mieux que les autres ? ##### Initialisation de la base de données

```
read.csv2("geoduction-data2.csv", sep=";", header=TRUE, na.strings = "", encoding = "UTF-8")->bdd
#exists('bdd')
```

Khi-Deux

```
bddKhiDeux = bdd[, c('Académie', 'Effectif.Présents.série.L', 'Effectif.Présents.série.ES', 'Effectif.Présents.série.M')]
# Petit clean des datas (Antoine)
bddKhiDeux[is.na(bddKhiDeux)] <- 0

# Cette portion de code suppose que bdd est ordonné par nom d'académie. (Antoine)

tableKhiDeux <- data.frame()
```

```

indiceCourant = 1
aca <- bddKhiDeux[1,"Académie"]

for(i in 1:nrow(bddKhiDeux)){
  if(aca != bddKhiDeux[i,"Académie"]){
    indiceCourant = indiceCourant + 1
    aca <- bddKhiDeux[i,"Académie"]
  }
  if (length(rownames(tableKhiDeux)) != 0 && !is.na(tableKhiDeux[indiceCourant,"Académie"]) && bddKhiDeux[i,"Académie"] != aca){
    tableKhiDeux[indiceCourant,"ES"] <- tableKhiDeux[aca,"ES"] + bddKhiDeux[i,"Effectif.Présents.série.L"]
    tableKhiDeux[indiceCourant,"L"] <- tableKhiDeux[aca,"L"] + bddKhiDeux[i,"Effectif.Présents.série.L"]
    tableKhiDeux[indiceCourant,"S"] <- tableKhiDeux[aca,"S"] + bddKhiDeux[i,"Effectif.Présents.série.S"]
  } else {
    tableKhiDeux <- rbind(tableKhiDeux, data.frame(Académie = aca,ES = bddKhiDeux[i,"Effectif.Présents.série.L"],L = bddKhiDeux[i,"Effectif.Présents.série.L"],S = bddKhiDeux[i,"Effectif.Présents.série.S"]))
  }
}

print(tableKhiDeux)

```

##	Académie	ES	S	L
## 1	AIX-MARSEILLE	4509	7678	2311
## 2	AMIENS	2744	4651	1530
## 3	BESANCON	1721	3099	767
## 4	BORDEAUX	5003	8344	2632
## 5	CAEN	2360	3614	1353
## 6	CLERMONT-FERRAND	1797	2923	1118
## 7	CORSE	359	565	301
## 8	CRETEIL	7065	11182	3253
## 9	DIJON	2263	4053	1231
## 10	GRENOBLE	5856	9084	2491
## 11	GUADELOUPE	755	1248	570
## 12	GUYANE	328	438	274
## 13	LA REUNION	1300	2441	985
## 14	LILLE	6128	11026	2726
## 15	LIMOGES	830	1691	617
## 16	LYON	5374	8980	2149
## 17	MARTINIQUE	604	1021	418
## 18	MAYOTTE	577	408	395
## 19	MONTPELLIER	3633	6687	2268
## 20	NANCY-METZ	3349	6085	1578
## 21	NANTES	5841	9111	2921
## 22	NICE	3344	5504	1696
## 23	ORLEANS-TOURS	3756	6253	1904
## 24	PARIS	4556	7358	2535
## 25	POITIERS	2596	3913	1509
## 26	REIMS	1909	3448	955
## 27	RENNES	5707	8741	2426
## 28	ROUEN	2996	4805	1606
## 29	STRASBOURG	3102	5056	1124
## 30	TOULOUSE	4249	8152	2265
## 31	VERSAILLES	11720	17807	4720

On dispose ici de la table de départ pour calculer le Khi Deux. Cette table nous donne l'agrégation des effectifs par filière et par académie.

```

# Calcul de la table des Abstrait
abstraitKhiDeux <- tableKhiDeux

for(i in 1:nrow(abstraitKhiDeux)){
  abstraitKhiDeux$ES[i] = sum(tableKhiDeux$ES)*sum(tableKhiDeux[i,'ES'],tableKhiDeux[i,'S'],tableKhiDeux[i,'L'])
  abstraitKhiDeux$S[i] = sum(tableKhiDeux$S)*sum(tableKhiDeux[i,'ES'],tableKhiDeux[i,'S'],tableKhiDeux[i,'L'])
  abstraitKhiDeux$L[i] = sum(tableKhiDeux$L)*sum(tableKhiDeux[i,'ES'],tableKhiDeux[i,'S'],tableKhiDeux[i,'L'])
}

#Calcul de la table des écarts
ecartsKhiDeux <- tableKhiDeux

for(i in 1:nrow(abstraitKhiDeux)){
  ecartsKhiDeux$ES[i] = tableKhiDeux$ES[i] - abstraitKhiDeux$ES[i]
  ecartsKhiDeux$S[i] = tableKhiDeux$S[i] - abstraitKhiDeux$S[i]
  ecartsKhiDeux$L[i] = tableKhiDeux$L[i] - abstraitKhiDeux$L[i]
}

#Calcul de la table des contributions
contribKhiDeux <- tableKhiDeux
i=1
for(i in 1:nrow(abstraitKhiDeux)){
  contribKhiDeux$ES[i] = ecartsKhiDeux$ES[i]*ecartsKhiDeux$ES[i] / abstraitKhiDeux$ES[i]
  contribKhiDeux$S[i] = ecartsKhiDeux$S[i]*ecartsKhiDeux$S[i] / abstraitKhiDeux$S[i]
  contribKhiDeux$L[i] = ecartsKhiDeux$L[i]*ecartsKhiDeux$L[i] / abstraitKhiDeux$L[i]
}

print(contribKhiDeux)

```

##	Académie	ES	S	L
## 1	AIX-MARSEILLE	2.258210485	7.056569e-01	0.3631176
## 2	AMIENS	3.150555771	1.986713e-01	11.1330769
## 3	BESANCON	1.760279322	9.677770e+00	14.3858369
## 4	BORDEAUX	1.230219800	1.685375e-01	5.4101212
## 5	CAEN	0.377787557	1.367897e+01	34.5472304
## 6	CLERMONT-FERRAND	1.923209177	6.332123e+00	43.0948651
## 7	CORSE	2.404496402	9.361501e+00	60.6730302
## 8	CRETEIL	7.535248899	8.096613e-01	5.1044715
## 9	DIJON	7.853642509	2.247062e+00	1.5552036
## 10	GRENOBLE	17.573030493	3.834758e-01	23.3115699
## 11	GUADELOUPE	4.901726708	7.653678e+00	67.1922477
## 12	GUYANE	0.023175128	2.119144e+01	74.2972403
## 13	LA REUNION	27.440309990	5.814052e-01	78.1061081
## 14	LILLE	5.999835840	3.431562e+01	52.0066428
## 15	LIMOGES	28.290086932	1.230328e+00	30.6420672
## 16	LYON	2.989952960	1.209383e+01	77.5452107
## 17	MARTINIQUE	3.224003272	2.392065e+00	28.8958467
## 18	MAYOTTE	43.449851924	1.378283e+02	145.4695221
## 19	MONTPELLIER	34.300492976	1.071730e+00	41.4092052
## 20	NANCY-METZ	6.712826255	1.650792e+01	13.9422166
## 21	NANTES	4.311357601	7.437503e+00	4.1081607
## 22	NICE	0.026830616	1.291225e-01	0.7899129
## 23	ORLEANS-TOURS	0.285534967	2.802474e-03	0.4394440

## 24	PARIS	0.338804687	6.446688e+00	29.8354127
## 25	POITIERS	0.826191992	2.037636e+01	48.2748095
## 26	REIMS	4.834099205	5.678761e+00	1.5001534
## 27	RENNES	21.575404587	1.368208e+00	19.9556804
## 28	ROUEN	0.005711176	3.389475e+00	10.5838682
## 29	STRASBOURG	7.610441089	7.200998e+00	77.7877942
## 30	TOULOUSE	37.007187202	2.740241e+01	0.8256151
## 31	VERSAILLES	62.920279491	1.369277e+00	83.5206781

Ici nous avons appliqué les étapes successives permettant de calculer le Khi Deux. Soit la corrélation entre les deux variables qualitatives : Académies et Filière.

Sur la table des contributions (ci-dessus) on peut observer que certaines régions et séries ont une contribution fortes à rendre dépendante ces deux variables.

On peut constater que certaines académies correspondantes a des zones géographiques périphériques ont une influence forte sur le khi Deux. Mayotte, Corse et Guadeloupe notamment. Dans ces régions la répartition entre les filières est modifiées et on trouve notamment une plus grande proportion de personnes en filière Littéraire.

Certaines académies de métropoles ont également des comportements particulier, par exemple l'académie de Versailles a une proportion particulièrement forte de ES et faible de L. Les académies de Limoges, Montpellier et Strasbourg ont également des comportements qui s'écartent des standards.

On constate finalement que la proportion de filière L a une forte tendance à varier alors que les filières ES et S ont souvent une proportion stable l'une par rapport à l'autre (environs un peu moins de deux fois plus de S que de ES). Ainsi de nombreuses académies ont une proportion de L élevée (DOM-TOM etc..) ou faible (Lyon, Lille, Strasbourg etc..)

Finalement on calcule le score global de khi deux

```
khideux <- chisq.test(tableKhiDeux[,c('S','ES','L')])
print(khideux)
```

```
##
## Pearson's Chi-squared test
##
## data:  tableKhiDeux[, c("S", "ES", "L")]
## X-squared = 1789.1, df = 60, p-value < 2.2e-16
```

Cet indicateur nous permet de dire que la situation géographique est certainement fortement dépendante de la répartition entre les filières. En effet la probabilité que la situation géographique soit indépendante de la répartition dans les différentes filières est inférieure à 2.2e-16.

Régression

Est ce que le taux de réussite des élèves en terminale S s'explique par la localisation académique ?

Problématique

Une interrogation récurrente vis à vis de la réussite scolaire est de se demander si la situation géographique d'un étudiant tend à lui offrir des chances supplémentaires d'obtenir son baccalauréat.

Pour essayer de déterminer si l'académie a un rôle prédominant dans la réussite de l'élève nous allons chercher à connaître l'impact de l'académie sur le taux de réussite au bac S, mais nous nous interrogerons aussi sur l'influence émise par les taux de réussite au baccalauréat L. Ainsi, un environnement, crée par la jointure entre une situation géographique donnée et un taux de réussite dans une autre filière donné, a-t-il un fort impact sur la réussite d'un élève passant son baccalauréat scientifique ?

Ainsi, nous allons créer une matrice comportant l'académie, l'effectif présent en série scientifique, le taux brut de réussite dans cette même série et le taux dans la série L.

```
bddReg = bdd[, c('Académie','Effectif.Présents.série.S', 'Taux.Brut.de.réussite.série.S', 'Taux.Brut.de
```

Nous procédons ensuite au nettoyage de notre matrice en retirant les valeurs nulles et en transformant les taux à une forme $0 < x < 1$. De plus nous retirons les valeurs aberrantes, soit celles où il n'y a pas d'élève inscrit dans les filières étudiées.

```
#valeur non définies mise à 0
bddReg[is.na(bddReg)] <- 0
#transformation des taux
bddReg[3] <- bddReg[3]/100
bddReg[4] <- bddReg[4]/100

df=data.frame(bddReg[1],bddReg[2],bddReg[3], bddReg[4])
#suppression des données aberrantes
df<-df[(df$Effectif.Présents.série.S>0 & df$Taux.Brut.de.réussite.série.S>0 & df$Taux.Brut.de.réussite.série.L>0 & df$Taux.Brut.de.réussite.série.L<1)]
```

Pour mener une étude par académie nous devons agréger l'ensemble des établissements scolaire appartenant à la même académie.

```
#regroupement des effectifs par académie
regData = aggregate(df$Effectif.Présents.série.S, by=list(df$Académie), FUN=sum)
#moyenne de l'ensemble des taux de réussite des lycées par académie
regData = c(regData,aggregate(df$Taux.Brut.de.réussite.série.S, by=list(df$Académie), FUN=mean)[2])
regData = c(regData,aggregate(df$Taux.Brut.de.réussite.série.L, by=list(df$Académie), FUN=mean)[2])
```

Nous créons maintenant notre model linéaire et nous allons procéder à la régression.

```
#création du modèle
df = data.frame(regData[1], regData[2], regData[3], regData[4])
col_headings <- c('Académie','Effectif', 'TxRéussiteS', 'TxRéussiteL')
names(df) <- col_headings
model<-lm(df$TxRéussiteS~df$Effectif+df$TxRéussiteL, data = df)
#affichage des résultats
summary(model)
```

```
##
## Call:
## lm(formula = df$TxRéussiteS ~ df$Effectif + df$TxRéussiteL, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.066971 -0.009052  0.002738  0.011954  0.032609
##
## Coefficients:
```

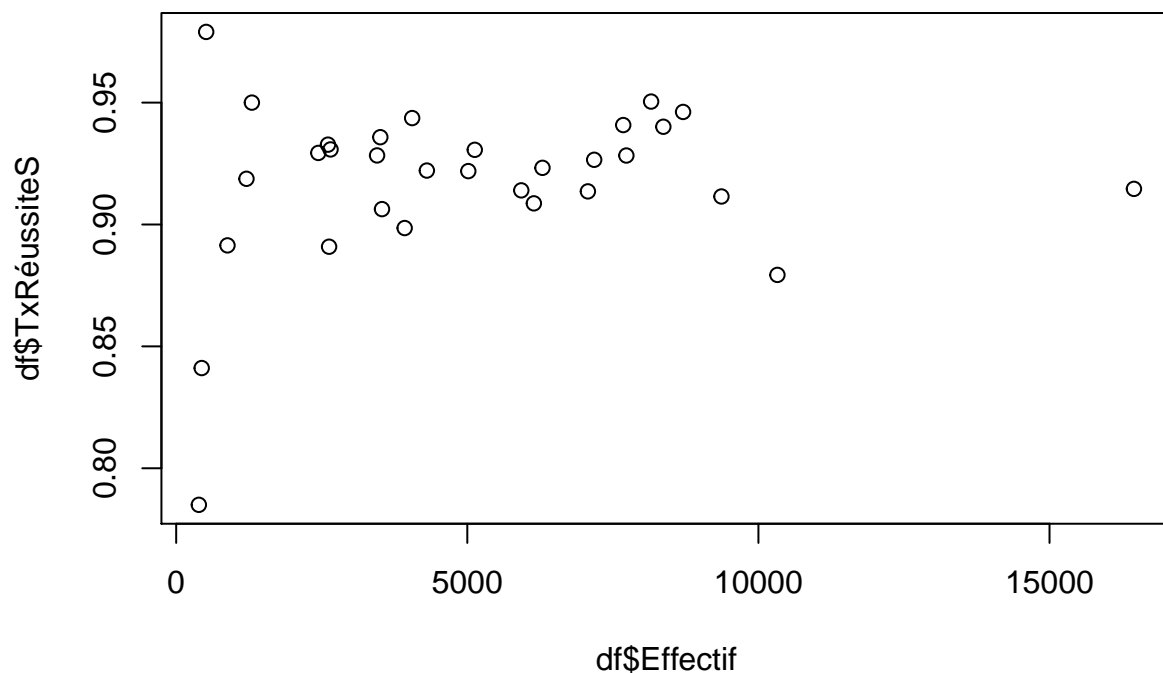
```
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)    4.434e-01  5.949e-02   7.453 4.06e-08 ***
## df$Effectif   -5.343e-08  1.049e-06  -0.051   0.96
## df$TxRéussiteL 5.170e-01  6.588e-02   7.847 1.51e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.01989 on 28 degrees of freedom
## Multiple R-squared:  0.699, Adjusted R-squared:  0.6775
## F-statistic: 32.51 on 2 and 28 DF,  p-value: 5.02e-08
```

Analyse

On constate que la valeur de R carré est élevée impliquant que le modèle a une importance sur le taux de réussite, selon une p-value extrêmement faible ($5.02e-08$) soit une précision de 1 sur 1 milliard. Cependant nous pouvons aussi observer que les deux variables utilisées n'ont pas le même impact sur notre résultat. En effet, l'effectif semble avoir un faible impact ($-5.343e-08$), tandis que le Taux de réussite en série L a un impact fort ($5.170e-01$).

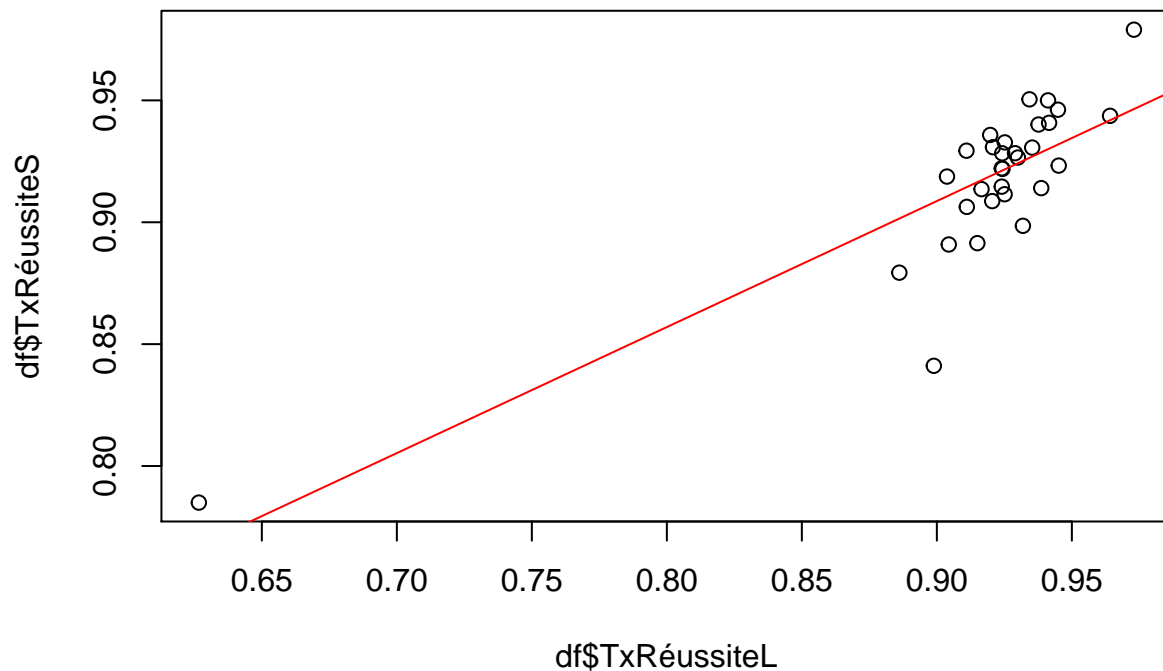
Nous pouvons représenter l'impact de l'effectif par le nuage de point suivant, et nous constatons que la droite de la fonction de régression

```
plot(df$Effectif,df$TxRéussiteS)
x <- seq(0,18000)
lines(x,x*-5.343e-08+4.434e-01,col="red")
```



Nous pouvons représenter l'impact du taux de réussite de la série L par le nuage de points suivant, et nous constatons que la droite de la fonction de régression montre une augmentation forte.

```
plot(df$TxRéussiteL,df$TxRéussiteS)
x <- seq(0,18000)
lines(x,x*5.170e-01+4.434e-01,col="red")
```



Conclusion

Ainsi, selon notre étude de donnée nous pouvons affirmer que l'environnement a un impact sur la réussite d'un élève, mais ce n'est pas la localisation qui crée cette empreinte mais la réussite des paires dans une série différente est-elle un facteur déterminant. On peut ainsi estimer qu'un environnement où une filière a un taux de réussite élevé impactera de manière positive les chances de réussite d'un élève.

Est ce que le taux de réussite des élèves en terminale S s'explique par la localisation des communes ?

Nous cherchons à savoir ici sur le fait qu'un étudiant inscrit au baccalauréat d'une commune a plus de chance de réussir que dans une autre commune. Nous allons regarder spécialement la série scientifique.

La première étape de ce cas de test réunie dans une nouvelle matrice, les colonnes "Ville", "Effectif.Présents.série.S" et "Taux.Brut.de.réussite.série.S". Nous allons tenter d'expliquer par la suite le taux de réussite de chaque ville par les effectifs inscrit dans ces mêmes localisations.


```
bddReg = bdd[, c('Ville', 'Effectif.Présents.série.S', 'Taux.Brut.de.réussite.série.S', 'Taux.Brut.de.ré.
```

Maintenant que nous disposons des données propres à l'étude de ce cas, nous avons besoin de nettoyer les données. Il faut notamment mettre des valeurs nulles dans les champs non remplis et ramener le taux à des valeurs comprises entre 0 et 1.

```
# Permet de mettre 0 dans les cases non remplies
bddReg[is.na(bddReg)] <- 0
# Ramène le pourcentage du taux de réussite à une valeur entre 0 et 1
bddReg[3] <- bddReg[3]/100
bddReg[4] <- bddReg[4]/100
```

Pour pallier à des villes où aucun candidat ne serait inscrit dans la série S, nous supprimons volontairement ces enregistrements qui sont considérés comme des individus aberrants pour notre étude. C'est ce que fait la portion de code suivante.

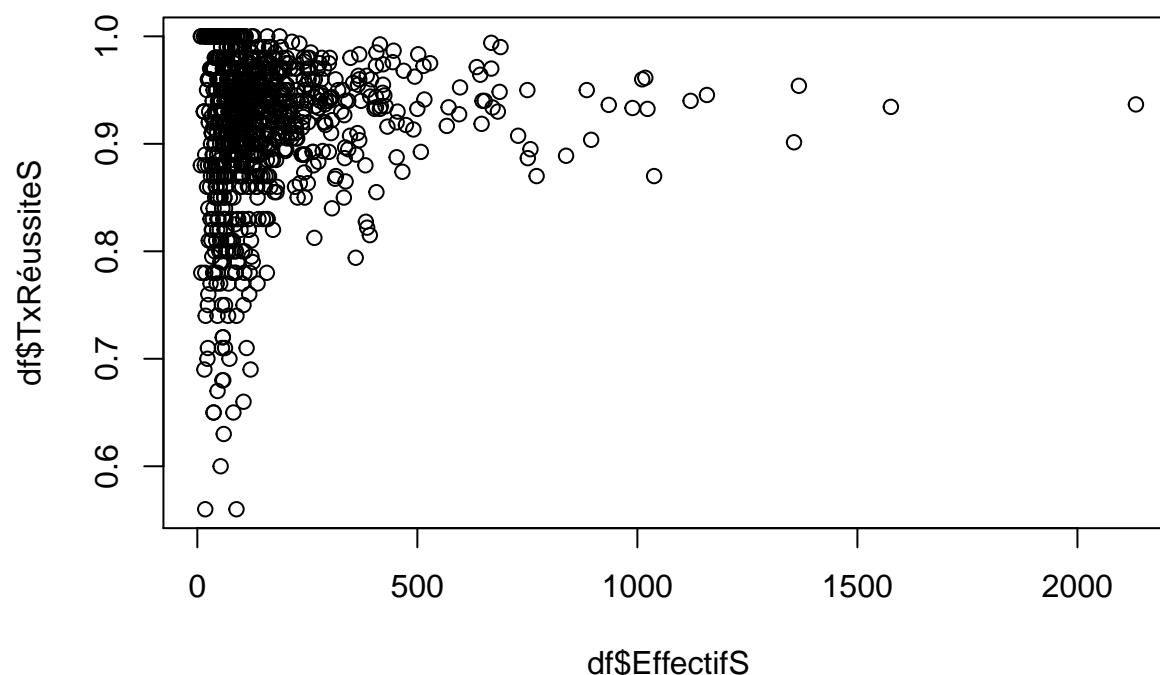
```
df=data.frame(bddReg[1],bddReg[2],bddReg[3], bddReg[4])
df<-df[(df$Effectif.Présents.série.S>0 & df$Taux.Brut.de.réussite.série.S>0 & df$Taux.Brut.de.réussite.série.L>0)]
```

Ce que nous allons faire maintenant, c'est faire un groupement par ville en faisant la somme des effectifs et la moyenne des taux de réussite de chaque établissement pour avoir un seul enregistrement par ville.

```
# Addition des effectifs groupé par Ville
regData = aggregate(df$Effectif.Présents.série.S, by=list(df$Ville), FUN=sum)
# Moyenne des taux de réussite des séries S assimilée
regData = c(regData,aggregate(df$Taux.Brut.de.réussite.série.S, by=list(df$Ville), FUN=mean)[2])
# Moyenne des taux de réussite des séries L assimilée
regData = c(regData,aggregate(df$Taux.Brut.de.réussite.série.L, by=list(df$Ville), FUN=mean)[2])
```

On peut donc maintenant s'apercevoir de la répartition du taux de réussite S en fonction des effectifs par ville.

```
df = data.frame(regData[1], regData[2], regData[3], regData[4])
col_headings <- c('Ville', 'EffectifS', 'TxRéussiteS', 'TxRéussiteL')
names(df) <- col_headings
plot(df$EffectifS,df$TxRéussiteS)
```



Nous allons donc effectuer une régression linéaire sur ces données pour tenter d'expliquer le taux de réussites par le lieu d'inscription du candidat au baccalauréat.

```
model<-lm(df$TxRéussiteS~df$EffectifS + df$TxRéussiteL, data = df)
summary(model)
```

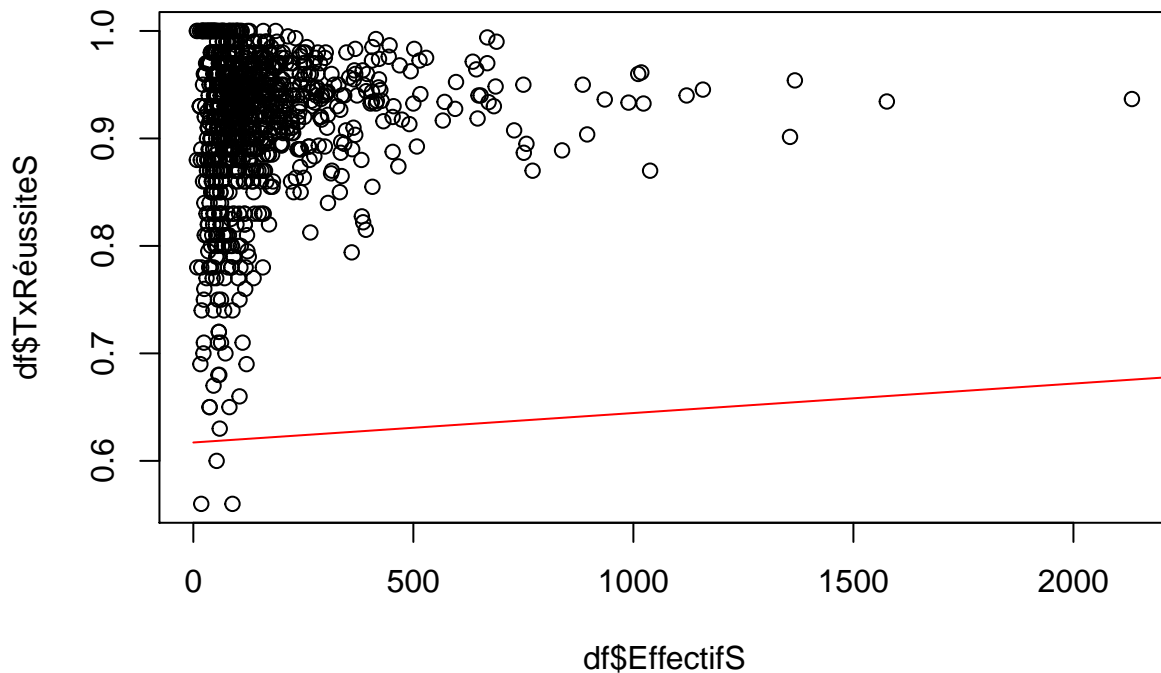
```
##
## Call:
## lm(formula = df$TxRéussiteS ~ df$EffectifS + df$TxRéussiteL,
##     data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.33674 -0.02687  0.00833  0.04207  0.14147
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   6.171e-01  2.256e-02  27.356   <2e-16 ***
## df$EffectifS   2.369e-05  1.069e-05   2.216   0.0269 *
## df$TxRéussiteL 3.228e-01  2.460e-02  13.118   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.06015 on 1014 degrees of freedom
## Multiple R-squared:  0.155, Adjusted R-squared:  0.1533
## F-statistic: 92.98 on 2 and 1014 DF,  p-value: < 2.2e-16
```

On s'aperçoit que l'effectif explique peu le taux de réussite. En effet, pour une unité du taux de réussite, l'effectif change de $2.741e-05$ ce qui est très petit.

Le R carré ajusté en tendant vers 0 (adjusted R square = 0.005677) nous indique aussi que l'effectif explique faiblement le taux de réussite avec environ 7 chance sur 1000 de se tromper donc cette prédiction est plutôt forte (p-value = 0.007613).

En traçant la droite $ax + b$ correspondant au modèle ($2.741e-05x + 6.171e-01$), on remarque sa faible pente et sa représentation plutôt horizontale ce qui indique aussi par le visuel un faible lien.

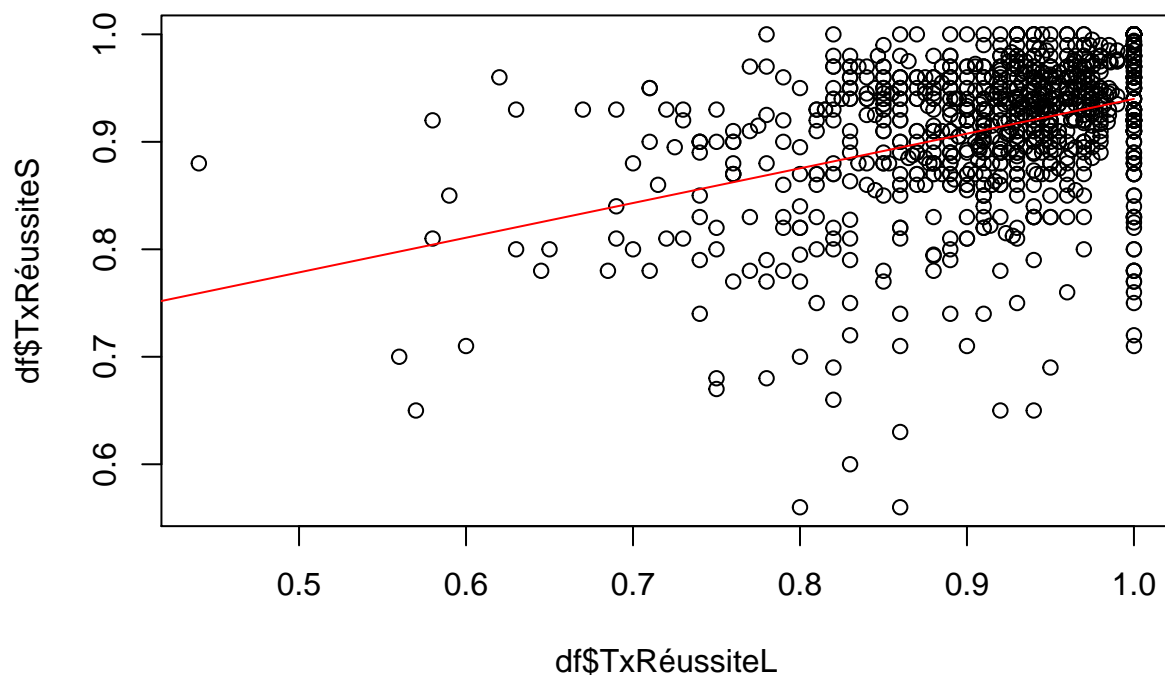
```
plot(df$EffectifS,df$TxRéussiteS)
x <- seq(0,2300)
lines(x,x*2.741e-05+6.171e-01,col="red")
```



En revanche pour le taux de réussite des séries L explique encore une fois le taux de réussite des séries S. On peut imaginer que le niveau général d'une ville irradie de L en S. Ce lien existe avec un R carré ajusté de 0.1533 avec une p-value infinitésimalement petite. Il y a donc quasiment aucune chance de se tromper sur ce point.

A titre informatif voici la droite du modèle qui représente ce lien :

```
plot(df$TxRéussiteL,df$TxRéussiteS)
x <- seq(0,1)
lines(x,x*3.228e-01+6.171e-01,col="red")
```



Regression simple

La régression linéaire simple est une méthode statistique permettant de modéliser la relation linéaire entre 2 variables quantitatives. Nous effectuerons notre régression linéaire sur 2 variables : - le taux de réussite (variable à expliquer) - les effectifs

Nous souhaitons analyser la relation entre le taux de réussite d'un établissement au baccalauréat (exprimé en pourcentage) et son effectif.

```
# [1] Extraction des champs qui nous intéressent
chu_reg_simple_extrac_data = bdd_complet[, c('Etablissement', 'Code.Etablissement', 'Effectif.Présents.Total.séries')]

# [2] Renommage des colonnes
names(chu_reg_simple_extrac_data)[names(chu_reg_simple_extrac_data)=="Etablissement"] <- "Nom_lycee"
names(chu_reg_simple_extrac_data)[names(chu_reg_simple_extrac_data)=="Code.Etablissement"] <- "Code_lycee"
names(chu_reg_simple_extrac_data)[names(chu_reg_simple_extrac_data)=="Effectif.Présents.Total.séries"] <- "Effectif"
names(chu_reg_simple_extrac_data)[names(chu_reg_simple_extrac_data)=="Taux.Brut.de.réussite.Total.séries"] <- "Taux"

# [3] Copie de l'extraction pour travailler dessus
chu_reg_simple_bdd = chu_reg_simple_extrac_data

# [4] Nettoyage des données
# suppression des lignes contenant des NA
chu_reg_simple_bdd <- na.omit(chu_reg_simple_bdd)
# conversion pour la reconnaissance des variables QUANTIT
```

```
for (i in 3:ncol(chu_reg_simple_bdd)) {
  chu_reg_simple_bdd[, i] <- as.numeric(as.character(chu_reg_simple_bdd[, i]))
}
# visualisation générale des données
str(chu_reg_simple_bdd)
```

```
## 'data.frame': 2288 obs. of 4 variables:
## $ Nom_lycee : Factor w/ 1769 levels "CITE SCOLAIRE INTERNATIONALE",...: 108 73 990 1318 15 524 1
## $ Code_lycee : Factor w/ 2288 levels "0010006B","0010010F",...: 50 53 54 55 57 51 56 52 224 157 .
## $ Effectif_lycee: num 41 172 31 165 21 309 152 148 58 425 ...
## $ Taux_reussite : num 93 93 94 96 100 94 94 89 90 94 ...
```

```
summary(chu_reg_simple_bdd)
```

```
##                               Nom_lycee      Code_lycee
## LYCEE ST JOSEPH                : 17      0010006B: 1
## LYCEE NOTRE DAME                : 16      0010010F: 1
## LYCEE JEANNE D'ARC              : 15      0010013J: 1
## LYCEE SAINT JOSEPH              : 12      0010014K: 1
## LYCEE LEONARD DE VINCI (GENERAL ET TECHNO.): 10      0010016M: 1
## LYCEE FENELON                  : 9        0010032E: 1
## (Other)                        :2209      (Other) :2282
## Effectif_lycee  Taux_reussite
## Min.   : 20.0    Min.   : 46.00
## 1st Qu.:104.0    1st Qu.: 90.00
## Median :186.0    Median : 94.00
## Mean   :201.4    Mean   : 92.49
## 3rd Qu.:283.0    3rd Qu.: 97.00
## Max.   :800.0    Max.   :100.00
##
```

```
plot(Taux_reussite ~ Effectif_lycee, data=chu_reg_simple_bdd)
```

```
# [5] Estimation du paramètre Effectif_lycee
# regression linéaire : nuage de points + droite de regression
chu_reg_simple <- lm(Taux_reussite ~ Effectif_lycee, data=chu_reg_simple_bdd)
summary(chu_reg_simple)
```

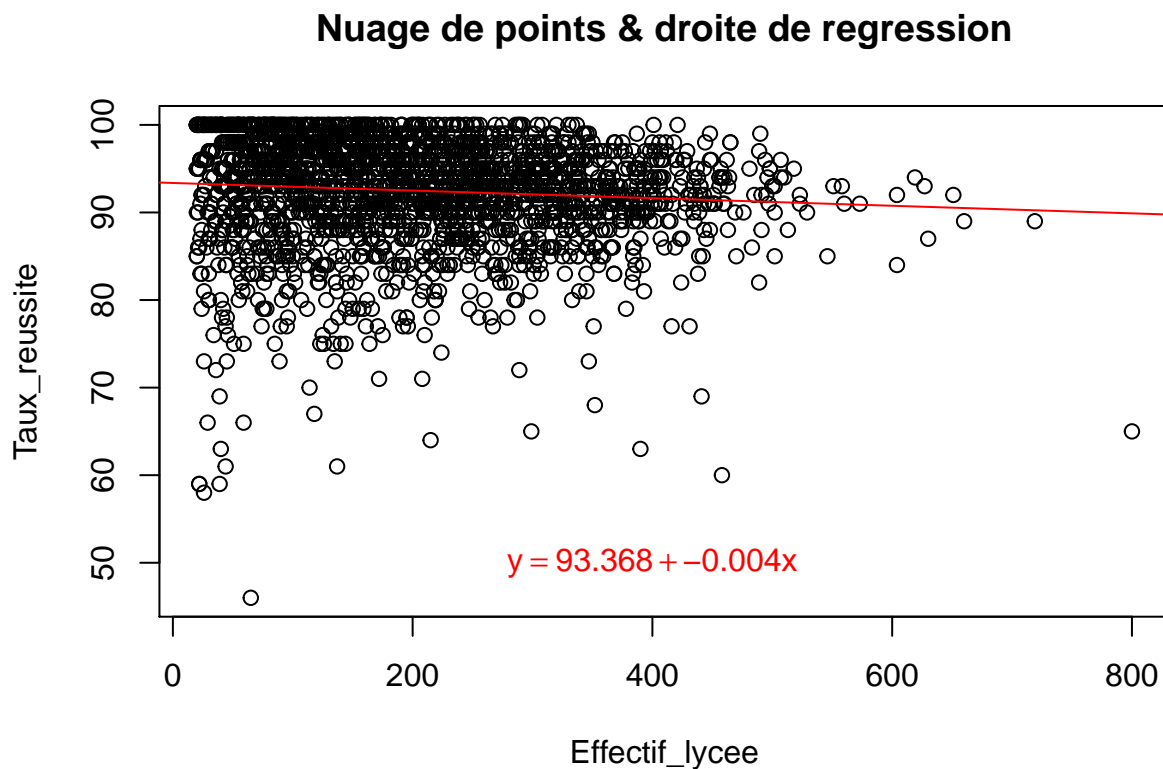
```
##
## Call:
## lm(formula = Taux_reussite ~ Effectif_lycee, data = chu_reg_simple_bdd)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -47.086  -2.692   1.110   4.354   8.458
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  93.367616   0.251276 371.574 < 2e-16 ***
## Effectif_lycee -0.004336   0.001074  -4.039 5.56e-05 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.122 on 2286 degrees of freedom
## Multiple R-squared:  0.007084,    Adjusted R-squared:  0.00665
## F-statistic: 16.31 on 1 and 2286 DF,  p-value: 5.555e-05

chu_reg_simple

##
## Call:
## lm(formula = Taux_reussite ~ Effectif_lycee, data = chu_reg_simple_bdd)
##
## Coefficients:
##      (Intercept)    Effectif_lycee
##           93.367616          -0.004336

abline(chu_reg_simple, col = 'red')
# affichage de l'équation de la droite de régression
title("Nuage de points & droite de regression")
text(400, 50, as.expression(substitute(y==b+a*x, list(
  a=round(chu_reg_simple$coefficients[2],3),
  b=round(chu_reg_simple$coefficients[1],3)
))), col = 'red')
```



Chaque point représente pour un lycée donné, l'effectif de l'établissement et son taux de réussite au baccalauréat. On remarque que la liaison entre les effectifs et le taux de réussite ne semble pas être linéaire.

Rédaction à revoir et améliorer

Std.Error = 0.251276 : Ecart-type associé à l'estimation des effectifs est petite, ce qui signifie un gage de stabilité du modèle et donc du pouvoir prédictif (valeur de b stable) Residual standard error = 6.122 : Estimateur de l'écart-type résiduel est faible donc bon pouvoir prédictif mais DDL à 2286 Multiple R-squared = 0.007084 : coeff de corrélation (% de variations expliquées par le modèle), R^2 doit être proche de 1 pour bon pouvoir explicatif

Le modèle n'a pas un bon pouvoir explicatif sur les données : $R^2=0.007084$ Le pouvoir prédictif risque d'être entaché par l'instabilité du coefficient b et une variance résiduelle importante

Regression multiple

```
# [1] Extraction des champs qui nous interessent
chu_reg_multi_extrac_data = bdd_complet[, c(
  'Secteur.Public.PU.Privé.PR',
  'Structure.p.dagogique.en.7.groupes',
  'Effectif.Présents.série.L',
  'Effectif.Présents.série.ES',
  'Effectif.Présents.série.S',
  'Taux.Brut.de.réussite.série.L',
  'Taux.Brut.de.réussite.série.ES',
  'Taux.Brut.de.réussite.série.S',
  'Taux.Brut.de.réussite.Total.séries',
  'Taux.accés.Brut.premi.re.BAC',
  'Taux.accés.Brut.terminale.BAC'
)]

# [2] Renommage des colonnes
names(chu_reg_multi_extrac_data) <- c(
  'Secteur_lycee',
  'Structure_lycee',
  'Effectif_L',
  'Effectif_ES',
  'Effectif_S',
  'Reussite_L',
  'Reussite_ES',
  'Reussite_S',
  'Reussite_Total',
  'Acces_prem_BAC',
  'Acces_term_BAC'
)

# [3] Copie de l'extraction pour travailler dessus
chu_reg_multi_bdd = chu_reg_multi_extrac_data
str(chu_reg_multi_bdd)
```

```
## 'data.frame': 2288 obs. of 11 variables:
## $ Secteur_lycee : Factor w/ 2 levels "PR","PU": 2 2 1 2 2 2 2 2 1 2 ...
## $ Structure_lycee: Factor w/ 7 levels "A","B","C","D",...: 1 3 2 6 1 2 6 3 6 3 ...
## $ Effectif_L : int 7 40 4 NA 4 54 NA 13 NA 35 ...
## $ Effectif_ES : int 14 33 4 NA 8 94 NA 37 NA 98 ...
```

```
## $ Effectif_S : int 20 39 12 82 9 79 97 57 NA 146 ...
## $ Reussite_L : int 100 100 100 NA 100 94 NA 62 NA 86 ...
## $ Reussite_ES : int 79 94 75 NA 100 93 NA 81 NA 93 ...
## $ Reussite_S : int 100 85 92 98 100 94 93 96 NA 95 ...
## $ Reussite_Total: int 93 93 94 96 100 94 94 89 90 94 ...
## $ Acces_prem_BAC: int 86 92 79 93 59 90 92 89 87 89 ...
## $ Acces_term_BAC: int 95 95 91 98 64 97 96 97 93 93 ...

# [4] Nettoyage des données
# on ne s'intéresse que aux données des établissements avec uniquement des filières généraux
chu_reg_multi_bdd <- subset(chu_reg_multi_bdd, Structure_lycee == "A")
# suppression des lignes contenant des NA
chu_reg_multi_bdd <- na.omit(chu_reg_multi_bdd)
# suppression des colonnes non utilisées pour la régression
chu_reg_multi_bdd <- chu_reg_multi_bdd[,-2]
# conversion pour la reconnaissance des variables QUANTI
for (i in 2:ncol(chu_reg_multi_bdd)) {
  chu_reg_multi_bdd[, i] <- as.numeric(as.character(chu_reg_multi_bdd[, i]))
}
# remplacement de la variable quali en quanti
chu_reg_multi_bdd[, 1] <- as.character(chu_reg_multi_bdd[, 1])
chu_reg_multi_bdd$Secteur_lycee[chu_reg_multi_bdd$Secteur_lycee=="PR"]<-"1"
chu_reg_multi_bdd$Secteur_lycee[chu_reg_multi_bdd$Secteur_lycee=="PU"]<-"0"
chu_reg_multi_bdd[, 1] <- as.numeric(chu_reg_multi_bdd[, 1])

# [5] Estimation des paramètres explicatives
chu_reg_multi <- lm(Reussite_Total ~ ., data=chu_reg_multi_bdd)
summary(chu_reg_multi)

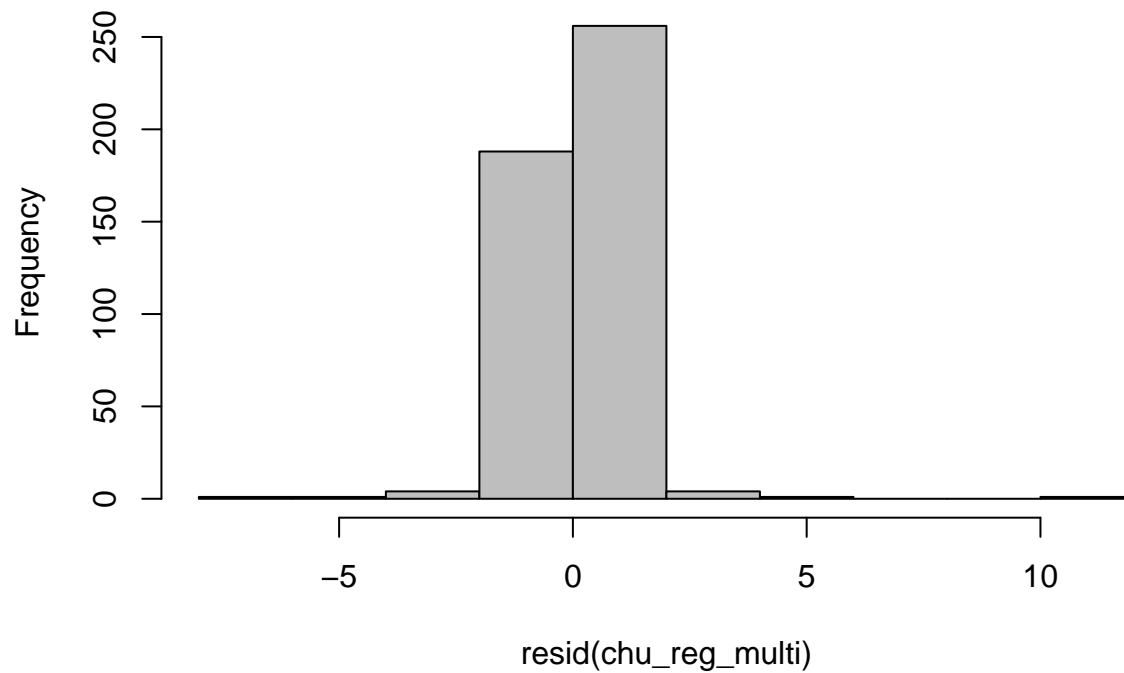
##
## Call:
## lm(formula = Reussite_Total ~ ., data = chu_reg_multi_bdd)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -6.9523 -0.3602  0.0774  0.3190 11.6003
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   3.2451546  0.9050005   3.586 0.000373 ***
## Secteur_lycee  0.1366615  0.1252899   1.091 0.275966
## Effectif_L     0.0002972  0.0037421   0.079 0.936726
## Effectif_ES   -0.0013847  0.0031227  -0.443 0.657658
## Effectif_S     0.0023884  0.0017156   1.392 0.164570
## Reussite_L     0.1822526  0.0067955  26.820 < 2e-16 ***
## Reussite_ES    0.3302230  0.0086354  38.241 < 2e-16 ***
## Reussite_S     0.4281752  0.0106790  40.095 < 2e-16 ***
## Acces_prem_BAC -0.0101643  0.0114705  -0.886 0.376031
## Acces_term_BAC  0.0319655  0.0196579   1.626 0.104637
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9859 on 446 degrees of freedom
## Multiple R-squared:  0.9757, Adjusted R-squared:  0.9752
```



```
## F-statistic: 1992 on 9 and 446 DF, p-value: < 2.2e-16
```

$R^2 = 97\%$ donc le modèle est précis Nous avons 97% de la variance du taux de réussite qui peut être expliquée par les variations de ...

```
#chu_reg_multi  
hist(resid(chu_reg_multi), col="grey", main="")
```



Conclusion en cours de rédaction

Conclusion Générale