

Geoducation

Antoine Drouhin, Aurélien Garret, Cécile Hu, Lucas Morel

Introduction

Base de données

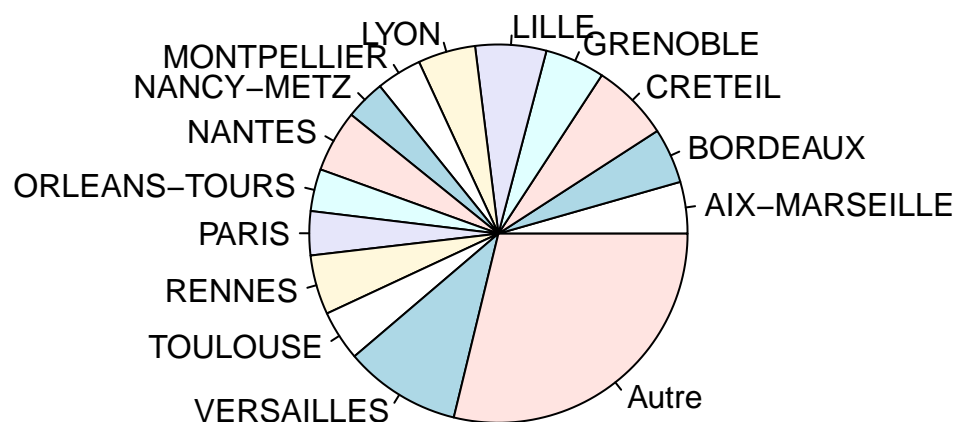
Notre base de données a été trouvée sur le site Data.gouv. Nous avons croisé deux jeux de données distincts. Le premier concerne des données sur la performance des lycées en France (taux de réussite etc). Le second présente des données géographiques pour l'ensemble des établissements scolaires français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

Répartition des effectifs par académie



NULL

L'intérêt et le questionnaire 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

Test de dépendance entre les académies et les filières (Méthode du khi-deux)

Dans cette partie nous allons tenter de déterminer s'il existe une relation entre la géographie (Académies) et la répartition des élèves entre les différentes filières générales du lycée. Nous ne savons pas si la répartition répond à une forme de quota de la part du ministère.

Nous avons choisi d'utiliser la méthode du khi-deux car cela répond particulièrement bien aux besoins de cette analyse de dépendance entre deux variables qualitatives (Des filières et des Académies). De plus nous avons choisi de montrer les différentes étapes du khi-deux, dans un souci de clarté mais également pour avoir une vision de l'impact de chaque combinaison académie/filière sur le résultat. Ce dernier point nous permettra de présenter une analyse plus fine de la dépendance entre les deux variables.

Notre hypothèse de départ est que la répartition des élèves entre les filières générales des lycées, d'une académie à l'autre, ne devrait pas changer.

Nous commençons donc par agréger les données de la base pour créer un tableau qui servira de point de départ à l'application de la méthode. Nous avons sommé les données les effectifs par académie. Nous avons également restreint notre étude aux colonnes qui nous intéressent, les effectifs présents dans les filières générales.

##	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

A partir de cette table nous allons calculer une table des valeurs théoriques que nous devrions trouver dans le tableau en fonction des effectifs totaux pour chaque filière et académie. (Voir Annexe1)

En fonction de ces deux tables on crée une table qui rend compte de l'influence des différentes combinaisons de variables sur le résultat du khi-deux. Cette table vas nous permettre de développer notre analyse ci-apres.

##	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

On peut observer que certaines régions et séries ont une contribution fortes à rendre dépendante ces deux variables.

On constate que certaines académies correspondantes à 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 métropole 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

##

```
## 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 2.2×10^{-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 ?

Déroulement du test

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.

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.

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

Nous créerons ensuite notre model linéaire et nous procéderons à la régression.

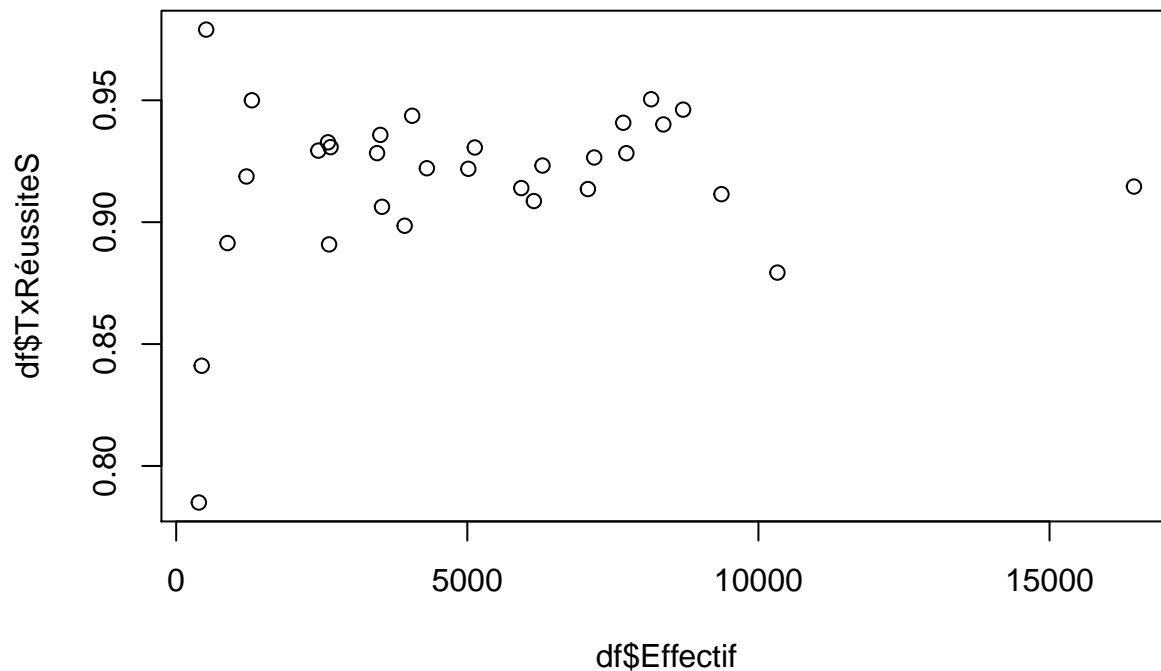
```
##
## Call:
## lm(formula = df$TxRéussiteS ~ df$Effectif + df$TxRéussiteL,
##     data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.066971 -0.009052  0.002738  0.011954  0.032609
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   4.434e-01  5.949e-02   7.453 4.06e-08 ***
## df$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
```

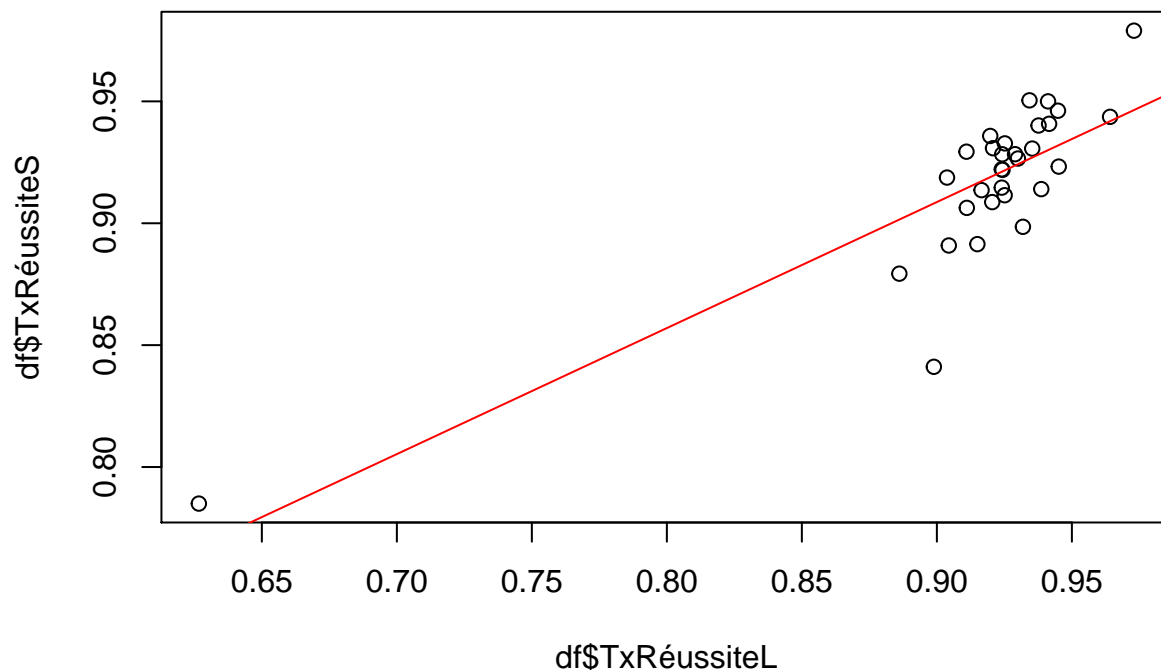
Analyse

Nous constatons 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 représente une augmentation extrêmement faible.



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.



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 situation géographique qui crée cette empreinte mais la réussite des paires dans une série différente est-elle un facteur déterminant. Nous pouvons ainsi conclure qu'un environnement où une filière a un taux de réussite élevé impactera de manière positive les chances de réussite d'un élève. Aussi, qu'un élève d'une grande métropole n'a concrètement pas plus ou moins de chances de réussir qu'un élève d'une région plus rurale.

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

Problématique

Nous cherchons à savoir ici si le fait qu'un étudiant inscrit au baccalauréat S d'une commune a plus de chance de réussir que dans une autre commune. Nous allons regarder spécialement la série scientifique. Nous voulons savoir aussi par la même occasion si le taux de réussite des séries L explique la réussite des séries S.

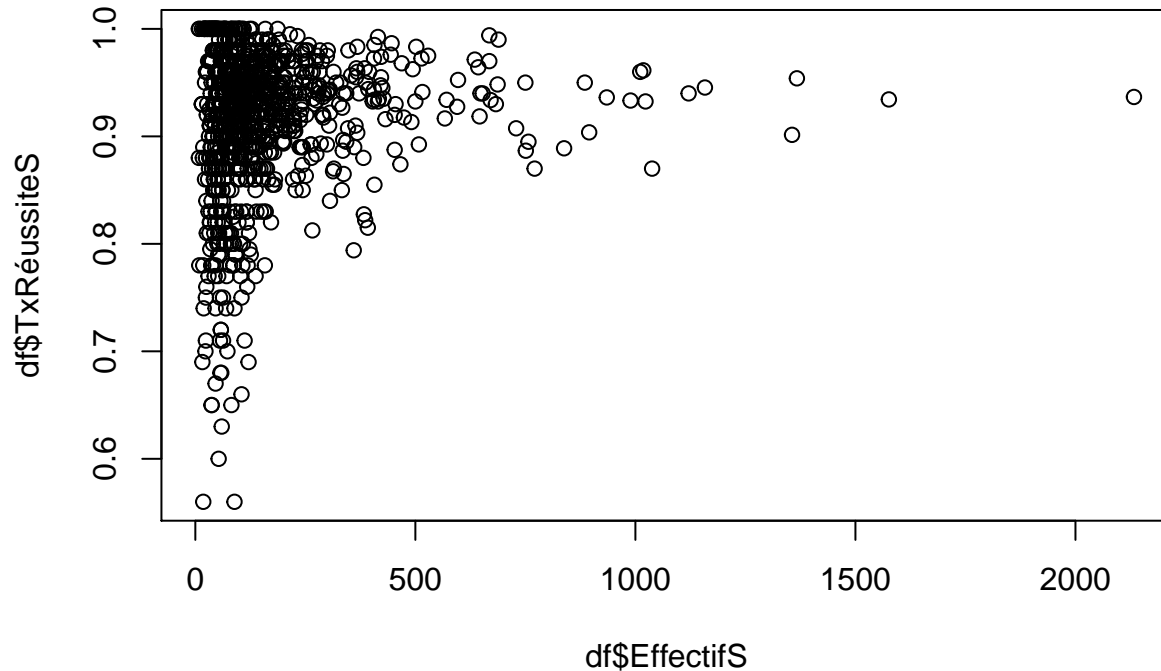
Déroulement du test

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. Nous irons aussi voir si le taux de réussite de la série L explique en partie le taux de réussite de la série S.

Pur parvenir à vérifier notre hypothèse, nous avons besoin de nettoyer les données aussi. Il faut notamment mettre des valeurs nulles dans les champs non remplis et ramener le taux à des valeurs comprises entre 0 et 1 pour correspondre au formalisme que nous avons choisi.

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.

Un groupement par ville est fait par la somme des effectifs et la moyenne des taux de réussite de chaque établissement pour avoir un seul enregistrement par ville.



Analyse des résultats

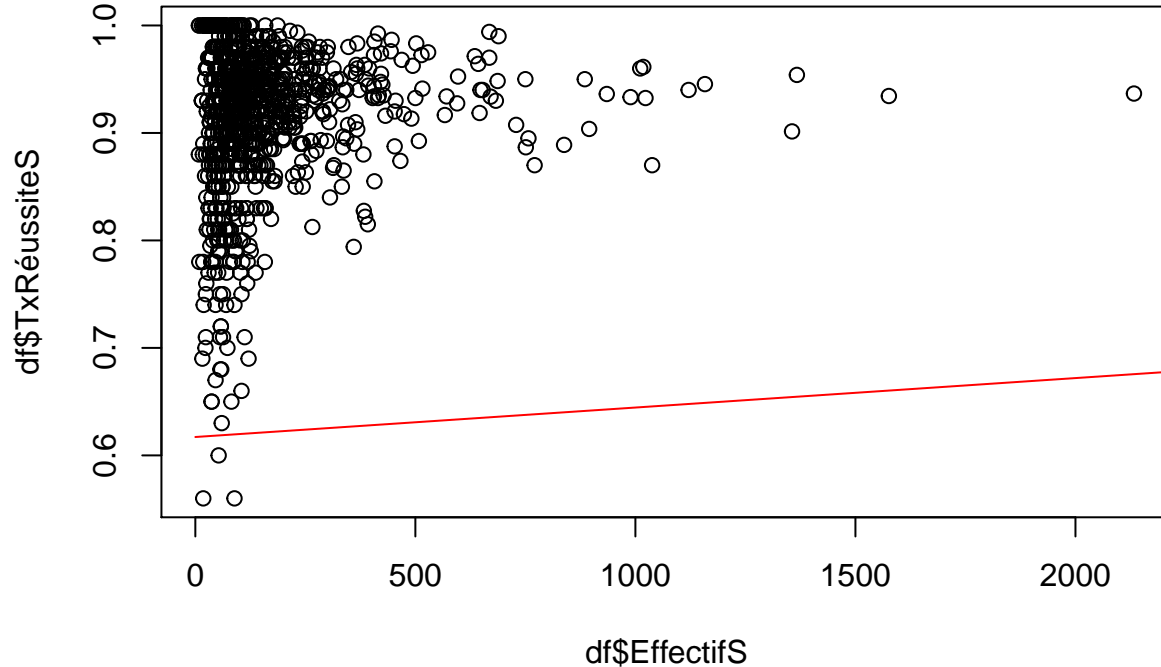
Une régression linéaire sur ces données est réalisée pour tenter d'expliquer le taux de réussites par le lieu d'inscription du candidat au baccalauréat S et du taux de réussite de ses camarades en série L.

```
##
## 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.

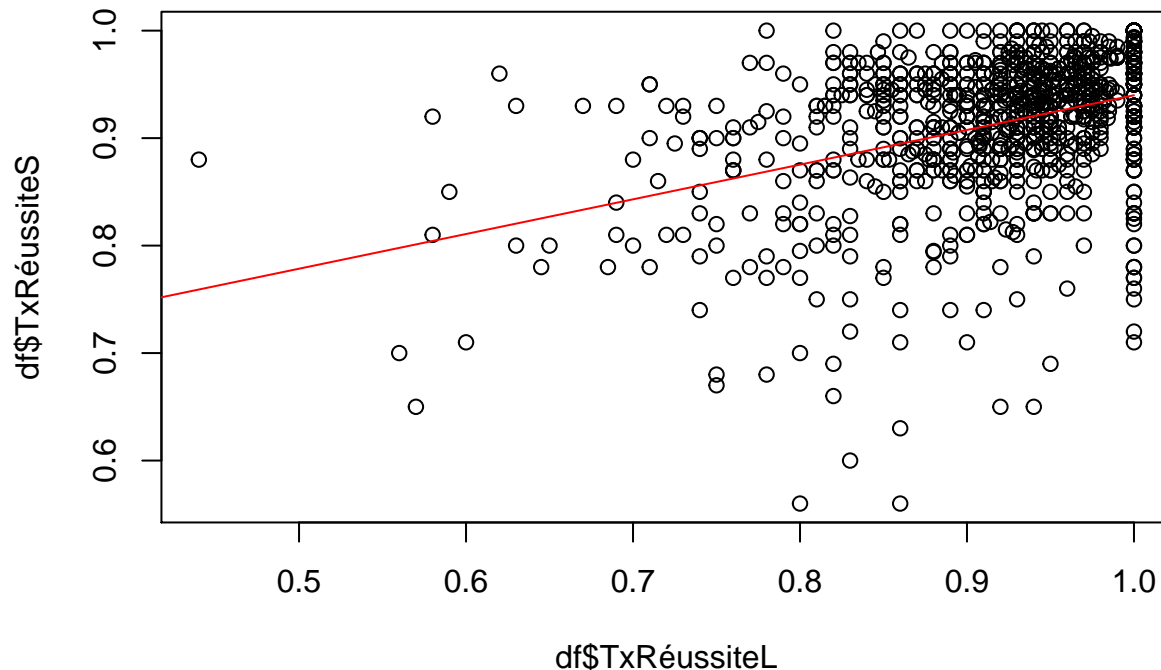
Le R carré ajusté en tendant vers 0 (adjusted R square = 0.005677) nous indique aussi que l'effectif explique faiblement le taux de réussite avec environ 7 chance sur 1000 de se tromper donc cette prédiction est plutôt forte (p-value = 0.007613).

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.



En revanche 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 infinitésimale petite. Il y a donc quasiment aucune chance de se tromper sur ce point.

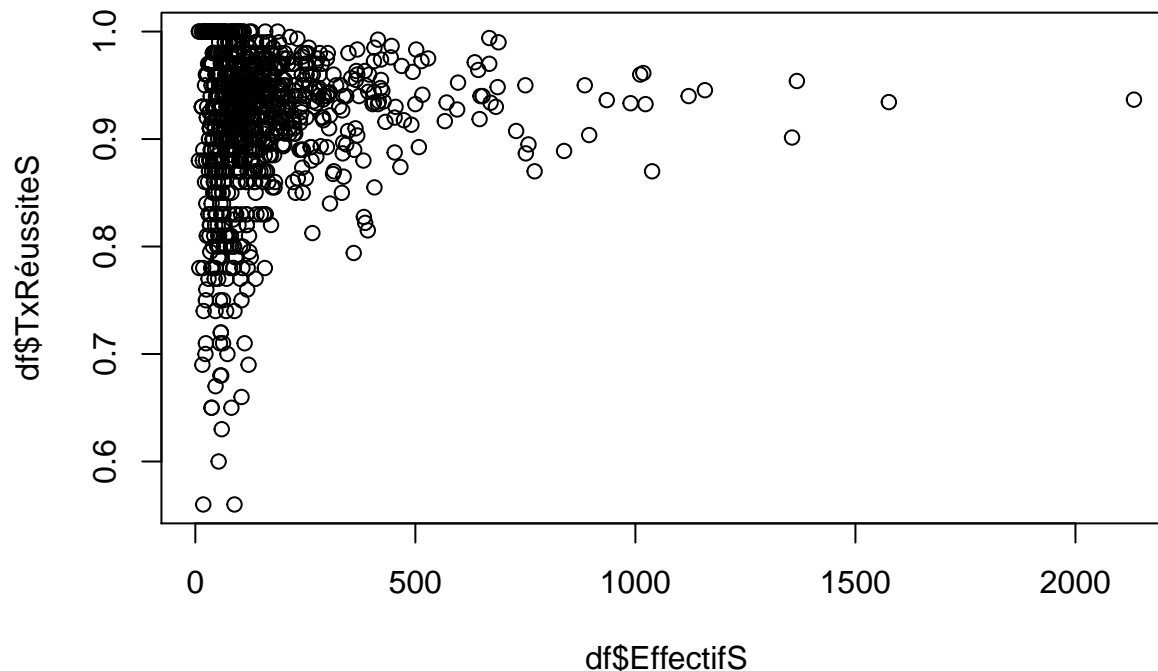
A titre informatif voici la droite du modèle qui représente ce lien :



Pour aller plus loin (vérification de l'indépendance de l'effectifs par ville et du taux de réussite des séries S)

Revenons un instant sur la repartition de notre nuage de point effectif serie S et taux de réussite série S. Il s'en dégage quelque chose de curieux que nous souhaitons éclaircir grace à test de chi deux.

Regardons à nouveau le nuage de points :



Nous pouvons voir qu'il semble que les effectifs les plus petits tendent à moins réussir pour certains tandis que nous ne pouvons affirmer cela pour les grands effectifs. Comme nous avons un effectif par ville on peut imaginer qu'il existe un lien entre le nombre d'inscrit d'une ville et sa réussite au baccalauréat serie S.

Pour parvenir à identifier cette dépendance nous allons mettre un place un test de chi deux. Nous allons dans un premier temps compartimenter les données comme ci dessous :

Catégorie effectifs * Inférieur à 300 : Petit * Entre 300 et 500 : Moyen * Entre 500 et 1000 : Grand

Catégorie taux de réussite * Entre 0.8 et 0.9 : Bien * Entre 0.9 et 1 : Très Bien

Le reste des données est écarté de notre test.

```
##          Effectif petit Effectif moyen Effectif grand
## Très bien          73557          19156          17365
## Bien              20996           5848           3625
```

```
##
## Pearson's Chi-squared test
##
## data:  chideuxData
## X-squared = 298.76, df = 2, p-value < 2.2e-16
```

Ce test d'indépendance par le khi deux nous démontre qu'il y a dépendance entre la taille de l'effectif et le taux de réussite en serie S. Un étudiant a donc plus de chance de réussir dans une grande ville. (P-value infiniment petite.)

Conclusion de la régression et du test d'indépendance

Notre régression linéaire n'a pas permis de démontrer une forte explication de la réussite au baccalauréat de chaque ville par ses effectifs ce qui se veut rassurant. Cependant nous avons vu que la réussite des séries L par ville expliquait en plus grande proportion la réussite des séries S. Alors pourquoi ? On peut imaginer que le niveau d'un établissement est uniforme indépendamment des séries qu'il propose et que cela se repercute sur le score de chaque ville.

Cependant, la représentation graphique du taux de réussite des séries S fonction de ses effectifs met en exergue certains échecs scolaires pour les petits effectifs et qui n'est pas présent dans les grands effectifs. Alors que peut on déduire de ça ? On pourrait imaginer qu'un petit groupe avec un élément perturbateur pourrait compromettre l'unification et la réussite de ce même groupe. Dans une grande masse de personne et par division de l'effectifs dans plusieurs établissements c'est peut être un fait qui tend à se lisser et qui est moins présent.

Regression multiple

Par quels paramètres la relation avec le taux de réussite des élèves inscrits dans un lycée de voie générale et technologique peut-il être expliqué ?

Problématique

Nous disposons d'une base de données présentant le taux de réussite à l'épreuve du baccalauréat en 2015 des différents établissements lycéens de France. Chaque établissement dispose d'une structure pédagogique différente. Un lycée proposant une large palette de formations est donc intéressant à étudier. Notre étude se portera sur les lycées appartenant au groupe C qui regroupe les lycées offrant des formations à la fois générales (séries S, ES et L) et technologiques (série STMG et autres séries). En effet, le taux de réussite pour ce type d'établissement est intéressant à analyser afin de déterminer les critères participants essentiellement à la variation de ce taux de réussite. Quelles sont ces paramètres qui entrent en jeu ? Sont-elles significatives ? À quelle proportion participent-elles au taux de réussite de l'établissement ? Quelle est la précision du modèle ?

Nous nous efforcerons de répondre à ces questions en analysant les données provenant de la base de l'éducation nationale. Nous présenterons dans un premier temps le déroulement de l'analyse, puis son explication, et finirons par une petite conclusion.

Déroulement de l'analyse

Il s'agit d'effectuer une régression linéaire multiple sur le taux de réussite des établissements du groupe C.

La régression linéaire multiple consiste, à partir de différentes variables quantitatives, d'expliquer la variable dite dépendante. On dit alors qu'on cherche à exprimer la variable à expliquer en fonction de plusieurs variables explicatives.

Le but de cette régression linéaire est de déterminer la significativité des variables, tout en minimisant le risque d'erreur pour prédire la précision du modèle.

Pour ce faire, nous allons procéder à une extraction des données qui nous intéressent afin d'effectuer notre régression. Nous allons sélectionner : - le secteur des établissements, - le groupe auquel il fait partie, - les effectifs pour chacune des filières générales de l'établissement, - le taux de réussite pour chacune des filières générales de l'établissement, - le taux de réussite de l'établissement, - le taux d'accès en première et en terminale

```

# [1] Extraction des champs qui nous intéressent
reg_multi_extrac_data = bdd[, c("Secteur.Public.PU.Privé.PR",
  "Structure.p.dagogique.en.7.groupe", "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.accès.Brut.premi.re.BAC", "Taux.accès.Brut.terminale.BAC")]

# [2] Renommage des colonnes
names(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
reg_multi_bdd = reg_multi_extrac_data
str(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 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 ...

```

Un nettoyage des données permet de ne travailler qu'avec des données cohérentes. On retirera alors les filières non-générales, les données non renseignées et nous remplacerons la variable qualitative (le secteur du lycée : public ou privé) en une variable quantitative (0 ou 1)

```

# [4] Nettoyage des données
reg_multi_bdd <- subset(reg_multi_bdd, Structure_lycee == "C") # établissements avec uniquement des fil

reg_multi_bdd <- na.omit(reg_multi_bdd) # suppression des lignes contenant des NA
reg_multi_bdd <- reg_multi_bdd[, -2] # suppression des colonnes non utilisées pour la régression
# conversion pour la reconnaissance des variables QUANTI
for (i in 2:ncol(reg_multi_bdd)) {
  reg_multi_bdd[, i] <- as.numeric(as.character(reg_multi_bdd[,
    i]))
}

# remplacement de la variable quali en quanti
reg_multi_bdd[, 1] <- as.character(reg_multi_bdd[, 1])
reg_multi_bdd$Secteur_lycee[reg_multi_bdd$Secteur_lycee == "PR"] <- "1"
reg_multi_bdd$Secteur_lycee[reg_multi_bdd$Secteur_lycee == "PU"] <- "0"
reg_multi_bdd[, 1] <- as.numeric(reg_multi_bdd[, 1])

```

Analyse

Afin d'estimer les paramètres, nous allons devoir écrire le modèle en utilisant la fonction `lm()` qui va permettre d'ajuster le modèle linéaire. Nos données contiennent 9 variables explicatives sur lesquelles nous appliquons

la régression linéaire afin de tenter d'expliquer le taux de réussite selon elles.

```
# [5] Estimation des paramètres explicatifs
reg_multi <- lm(Reussite_Total ~ ., data = reg_multi_bdd)
summary(reg_multi)

##
## Call:
## lm(formula = Reussite_Total ~ ., data = reg_multi_bdd)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -6.6563 -0.7052  0.1338  0.9493  6.9808
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    0.2871084   2.0577268    0.140   0.8891
## Secteur_lycee    0.9936119   0.2426447    4.095 5.04e-05 ***
## Effectif_L      0.0062482   0.0058215    1.073   0.2837
## Effectif_ES     -0.0009047   0.0049814   -0.182   0.8560
## Effectif_S      -0.0056130   0.0031790   -1.766   0.0782 .
## Reussite_L       0.1133238   0.0110557   10.250 < 2e-16 ***
## Reussite_ES      0.2039211   0.0153107   13.319 < 2e-16 ***
## Reussite_S       0.2606424   0.0150199   17.353 < 2e-16 ***
## Acces_prem_BAC   0.0314865   0.0325852    0.966   0.3344
## Acces_term_BAC   0.3801059   0.0486518    7.813 4.31e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.574 on 431 degrees of freedom
## Multiple R-squared:  0.9164, Adjusted R-squared:  0.9146
## F-statistic: 524.9 on 9 and 431 DF,  p-value: < 2.2e-16
```

Nous remarquons que 4 variables sont alors significatives pour le modèle : le secteur, le taux de réussite en L, ES et S, et le taux d'accès en terminale car la p-value est petite. Les coefficients estimés de ces paramètres étant petits, ils montrent la petite variation quand le taux de réussite de l'établissement augmente de 1%. Le R^2 ajusté mesure le pourcentage de la variance de la variable expliquée par la variance de toutes les variables explicatives. On remarque alors qu'elle est de 91% ce qui signifie que notre modèle est précis à 91%.

Conclusion

L'analyse des données des établissements du groupe C montre alors que le secteur du lycée (public ou privé) a un impact sur le taux de réussite. Les lycées privés ont un taux de réussite meilleur que les lycées publics (coefficient estimé positif). Les taux de réussites des filières générales ont également un impact sur le taux de réussite d'un établissement de groupe C. Le taux d'accès des élèves de terminale joue aussi sur le taux de réussite. Ce modèle explique donc avec une bonne précision (91%) le taux de réussite d'un établissement de groupe C en fonction de ces paramètres étudiés

Conclusion générale

Les différentes analyses effectuées sur notre base de données montrent qu'il existe un lien très fort entre la situation géographique des établissements lycéens et les filières proposées à l'enseignement. L'effectif d'un

établissement ne joue pas sur sa réussite au baccalauréat. En revanche, pour des petits effectifs, la réussite de l'établissement pourrait s'expliquer certaines filières en particulier. Dans les lycées polyvalents, le taux de réussite au baccalauréat s'explique beaucoup par la réussite des filières générales. Et il semblerait aussi que les établissements privés présentent un meilleur taux de réussite que les publics.

Annexes

Initialisation de la base de données

```
bdd <- read.csv2("geoducation-data2.csv", sep = ";", header = TRUE,
  na.strings = "", encoding = "UTF-8")
```

Annexes 1 - Test de dépendance entre Académie et Filière

```
bddKhiDeux = bdd[, c("Académie", "Effectif.Présents.série.L",
  "Effectif.Présents.série.ES", "Effectif.Présents.série.S")]
# 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"] == tableKhiDeux[indiceCourant,
    "Académie"]) {
    tableKhiDeux[indiceCourant, "ES"] <- tableKhiDeux[aca,
      "ES"] + bddKhiDeux[i, "Effectif.Présents.série.ES"]
    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.ES"],
      S = bddKhiDeux[i, "Effectif.Présents.série.S"],
      L = bddKhiDeux[i, "Effectif.Présents.série.L"]))
  }
}

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

```
# 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"])/sum(tableKhiDeux$ES,
    tableKhiDeux$S, tableKhiDeux$L)
  abstraitKhiDeux$S[i] = sum(tableKhiDeux$S) * sum(tableKhiDeux[i,
    "ES"], tableKhiDeux[i, "S"], tableKhiDeux[i, "L"])/sum(tableKhiDeux$ES,
    tableKhiDeux$S, tableKhiDeux$L)
  abstraitKhiDeux$L[i] = sum(tableKhiDeux$L) * sum(tableKhiDeux[i,
    "ES"], tableKhiDeux[i, "S"], tableKhiDeux[i, "L"])/sum(tableKhiDeux$ES,
    tableKhiDeux$S, tableKhiDeux$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]
}
```

```

    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

```

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
```

Annexes 2 - Expliquer le taux de réussite des séries S par les effectifs et le taux de réussite des séries L

Régression Taux de réussite des séries S ~ Effectifs des séries S par Académie + Taux de réussite des séries L

```
bddReg = bdd[, c("Académie", "Effectif.Présents.série.S",
  "Taux.Brut.de.réussite.série.S", "Taux.Brut.de.réussite.série.L")]

# 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), ]

# 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])

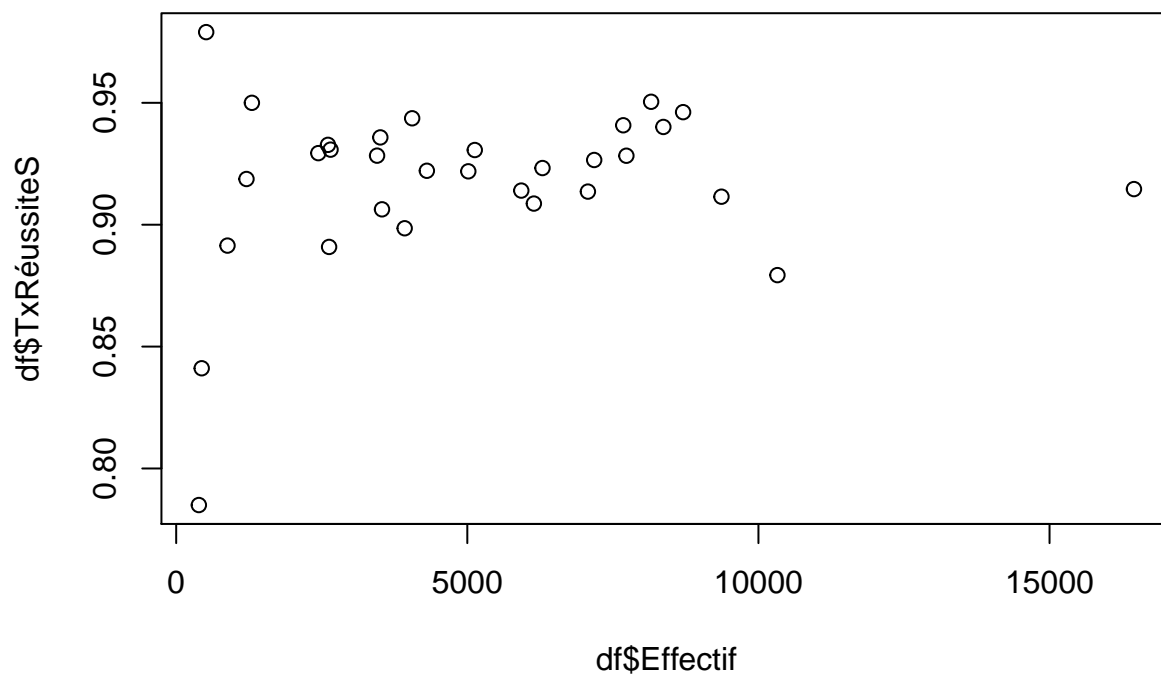
# création du modèle
df = data.frame(regData[1], regData[2], regData[3], regData[4])
col_headings <- c("Académie", "Effectif", "TxRéussiteS", "TxRéussiteL")
names(df) <- col_headings
model <- lm(df$TxRéussiteS ~ df$Effectif + df$TxRéussiteL,
  data = df)
# affichage des résultats
summary(model)
```

```
##
## 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
```

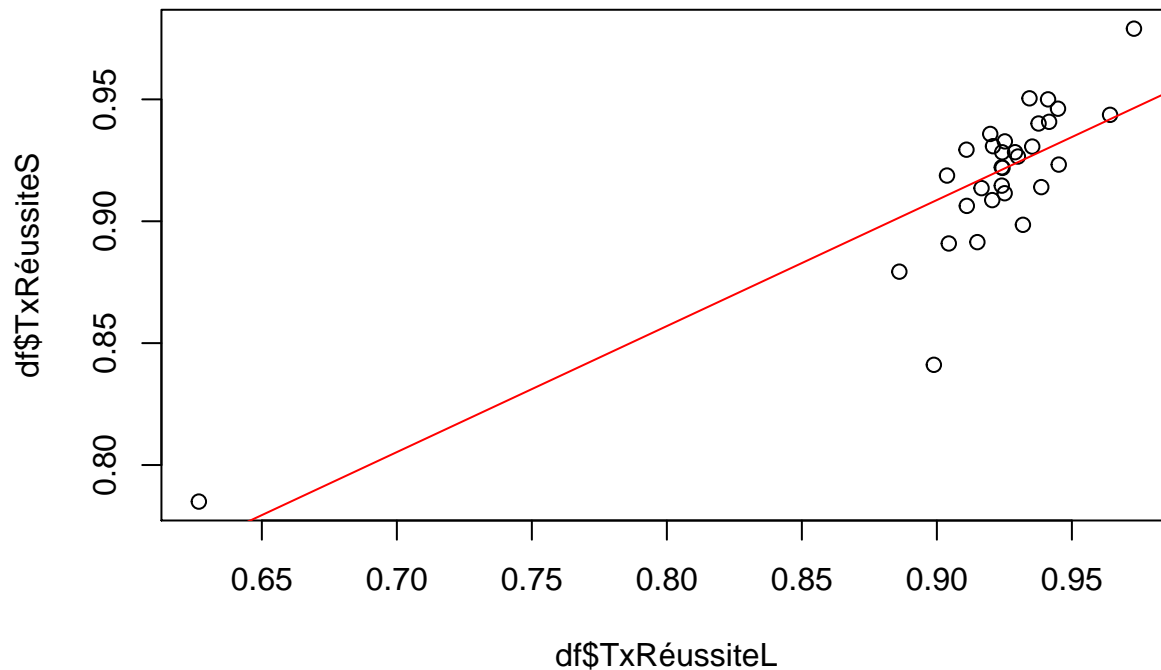


```
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   4.434e-01  5.949e-02   7.453 4.06e-08 ***
## df$Effectif  -5.343e-08  1.049e-06  -0.051   0.96
## df$TxRéussiteL 5.170e-01  6.588e-02   7.847 1.51e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 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
```

```
plot(df$Effectif, df$TxRéussiteS)
x <- seq(0, 18000)
lines(x, x * -5.343e-08 + 0.4434, col = "red")
```



```
plot(df$TxRéussiteL, df$TxRéussiteS)
x <- seq(0, 18000)
lines(x, x * 0.517 + 0.4434, col = "red")
```



Régression Taux de réussite des séries S ~ Effectifs des séries S par Ville + Taux de réussite des séries L + Test du chi-deux entre Taux de réussite S et Effectifs des séries S

```
bddReg = bdd[, c("Ville", "Effectif.Présents.série.S", "Taux.Brut.de.réussite.série.S",
  "Taux.Brut.de.réussite.série.L")]

# 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

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), ]

# 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])

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

```

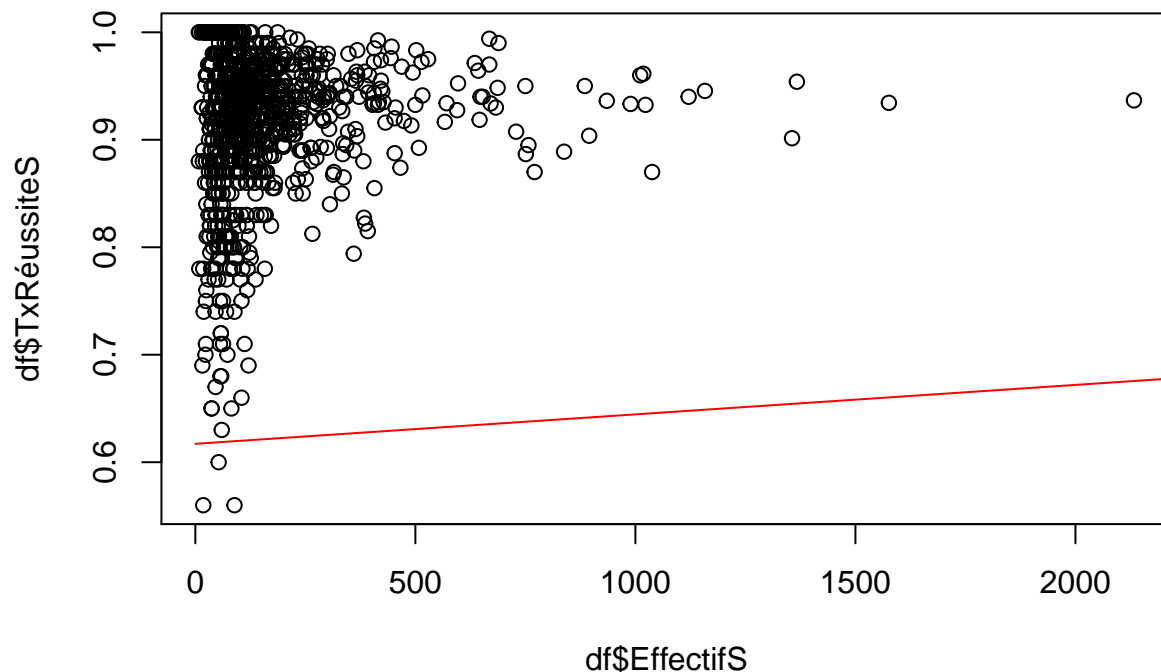
plot(df$EffectifS, df$TxRéussiteS)

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

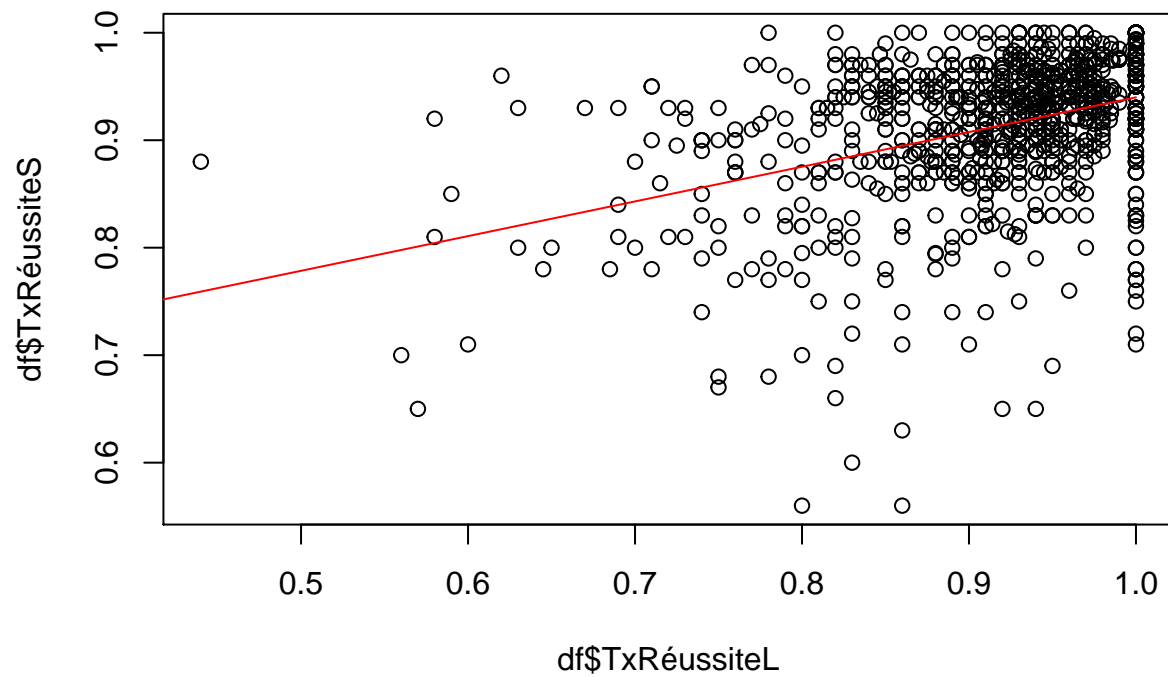
##
## Call:
## lm(formula = df$TxRéussiteS ~ df$EffectifS + df$TxRéussiteL,
##     data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.33674 -0.02687  0.00833  0.04207  0.14147
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  6.171e-01  2.256e-02  27.356  <2e-16 ***
## df$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

plot(df$EffectifS, df$TxRéussiteS)
x <- seq(0, 2300)
lines(x, x * 2.741e-05 + 0.6171, col = "red")

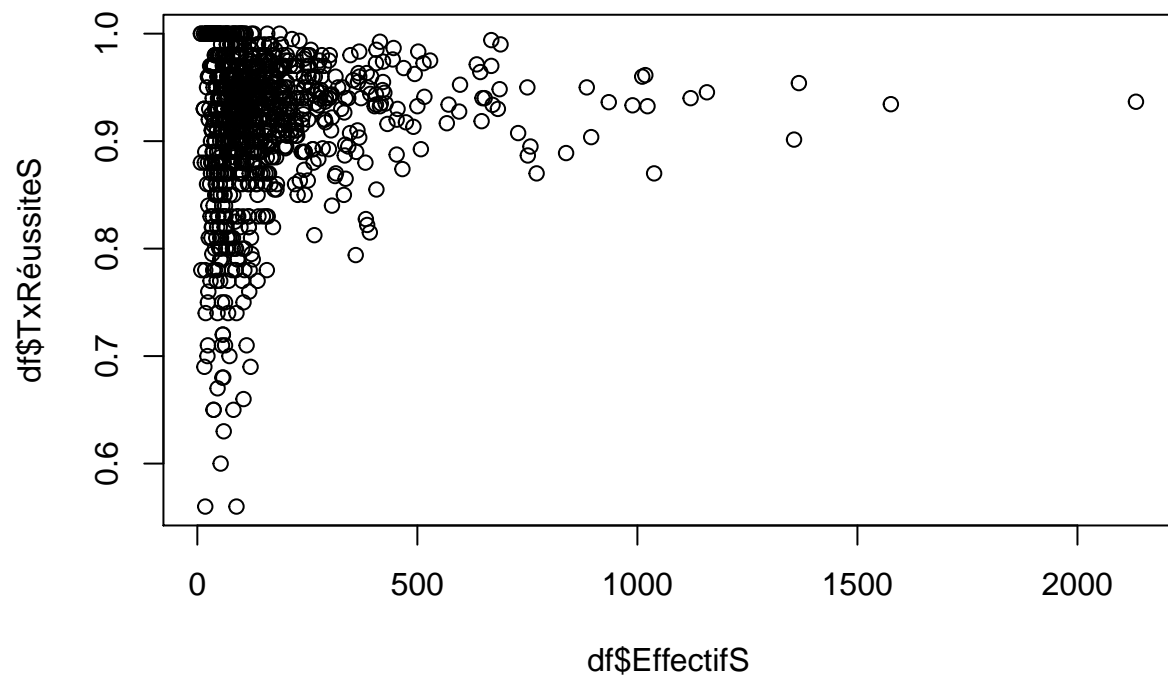
```



```
plot(df$TxRéussiteL, df$TxRéussiteS)
x <- seq(0, 1)
lines(x, x * 0.3228 + 0.6171, col = "red")
```



```
plot(df$EffectifS, df$TxRéussiteS)
```



```
# Petit - Très bien
PetitTBien <- sum(df$EffectifS[which(df$EffectifS >= 1 & df$EffectifS <
  300 & df$TxRéussiteS >= 0.9 & df$TxRéussiteS <= 1)])
```

```

# Petit - Bien
PetitBien <- sum(df$EffectifS[which(df$EffectifS >= 1 & df$EffectifS <
  300 & df$TxRéussiteS >= 0.8 & df$TxRéussiteS < 0.9)])

# Moyen - Très bien
MoyenTBien <- sum(df$EffectifS[which(df$EffectifS >= 300 & df$EffectifS <
  500 & df$TxRéussiteS >= 0.9 & df$TxRéussiteS <= 1)])

# Moyen - Bien
MoyenBien <- sum(df$EffectifS[which(df$EffectifS >= 300 & df$EffectifS <
  500 & df$TxRéussiteS >= 0.8 & df$TxRéussiteS < 0.9)])

# Grand - Très bien
GrandTBien <- sum(df$EffectifS[which(df$EffectifS >= 500 & df$EffectifS <=
  1000 & df$TxRéussiteS >= 0.9 & df$TxRéussiteS <= 1)])

# Grand - Bien
GrandBien <- sum(df$EffectifS[which(df$EffectifS >= 500 & df$EffectifS <=
  1000 & df$TxRéussiteS >= 0.8 & df$TxRéussiteS < 0.9)])

# Unification des données dans une matrice
chideuxData <- data.frame(c(PetitTBien, PetitBien), c(MoyenTBien,
  MoyenBien), c(GrandTBien, GrandBien))

# Nommage des colonnes et des lignes
names(chideuxData) <- c("Effectif petit", "Effectif moyen", "Effectif grand")
rownames(chideuxData) <- c("Très bien", "Bien")

# Affichage de nos effectifs réels
chideuxData

```

```

##           Effectif petit Effectif moyen Effectif grand
## Très bien           73557           19156           17365
## Bien                20996            5848            3625

```

```
chisq.test(chideuxData)
```

```

##
## Pearson's Chi-squared test
##
## data:  chideuxData
## X-squared = 298.76, df = 2, p-value < 2.2e-16

```

Annexes 3 - Régression multiple

```

## '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 ...

##
## Call:
## lm(formula = Reussite_Total ~ ., data = 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

```

