Juliette schlegel

université Sorbonne Paris Nord

M1 APE-MIF

Devoir d’econometrie

dans le cadre du cours de technique quantitative dispensé par Hugo Harari-Kermadec

# Résumé de l’article

Pour réaliser ce devoir, nous allons nous appuyer sur l’article de recherche Le passage à l’université : brassage ou ségrégation sociale ? Mesure de la polarisation du système universitaire français (2007-2015) réalisé par Hugo Harari-Kermadec, maître de conférences HDR au département d’économie et de gestion à l’ENS Paris-Saclay et Chercheur à l’IDHES et par R. Avouac. L’article sera publié en 2020 sans la revue *Economie et Statistique* éditée par l’INSEE.

D’après les auteurs, le nombre croissant de systèmes d’enseignement supérieurs à l’international couplé avec la facilité de déplacement des étudiants, pouvant ainsi intégrer des universités qui ne sont pas nationales, sont en grande partie à l’origine de l’augmentation d’indicateurs quantitatifs des performances des universités pour permettre de les classer à l’international. Ils indiquent par la suite un certain nombre de programmes nationaux et internationaux qui, s’appuyant sur des indicateurs quantitatifs, vont orienter les politiques publiques et ainsi les fonds destinés à l’enseignement supérieur publique et leurs priorités. Ces indicateurs quantitatifs comparent la performance des universités selon certains critères, orientant la direction des universités à améliorer justement ces critères, attirés par les fonds publics, peut-être au détriment de performances qualitatives plus difficilement mesurables. Par exemple, le classement de Shangaï, établi par des chercheurs de l'université Jiao-tong de Shanghai, est publié annuellement depuis 2003. Il classe les 500 universités du monde les plus performantes selon des critères spécifiques mesurant principalement la quantité et la qualité de la recherche produite par chaque université. On peut donc supposer que le choix de ces critères va orienter les politiques publiques sur la recherche afin de remonter dans ce classement très médiatisé sans pour autant améliorer la qualité de l’enseignement pour tous.

Les critiques adressées à ce type de classement sont avant tout méthodologique. La pertinence et les possibles biais des critères retenus pour juger de la qualité de l’enseignement dispensé dans les universités sont remis en cause. On peut citer comme exemple, toujours par rapport au classement de Shangaï, le nombre de prix Nobels et de médailles Fields remportés par les anciens élèves d’une université , qui avantagerai les plus anciennes de ces dernières sans pour autant garantir une qualité d’enseignement actuelle ; ou encore les problèmes que peut poser les récompenses attribuées à des recherches issues de partenariats.

Mais les critiques retenues par Mr. Harari-Kermadec et Mr Avouac sont de l’ordre de la répercussion de ces classements sur l’enseignement plutôt que de sa pertinence. La prise en compte de ses classements par les politiques publiques va redéfinir les priorités des directeurs d’université au-delà de leurs objectifs primaires de production de connaissance. Et au-delà de l’échelle locale, l’inquiétude énoncée dans l’article est celle d’une disparité entre les universités. D’une part, celles de « rang mondiale » devenant la priorité des politiques publiques et donc des ressources qui sont consacrées à l’enseignement public ; d’autre part, celles de « second rang » qui devront se partager les restes. Une bipolarité de l’accès aux connaissances existant déjà entre les universités publiques et les écoles privées souvent très chères. Pour appuyer cette critique, les auteurs rappellent que l’accès aux études supérieures en France par le biais des universités a, par le passé, permis une diminution des disparités dues aux origines sociales des étudiants.

L’objectif des recherches effectuées par Mr. Harari-Kermadec et Mr Avouac serait de savoir si les origines sociales des étudiants s’observent à travers une polarisation du système universitaire français. Et de savoir si cette disparité se serait accentuée depuis la parution d’indicateurs quantitatifs nationaux et internationaux des performances des universités.

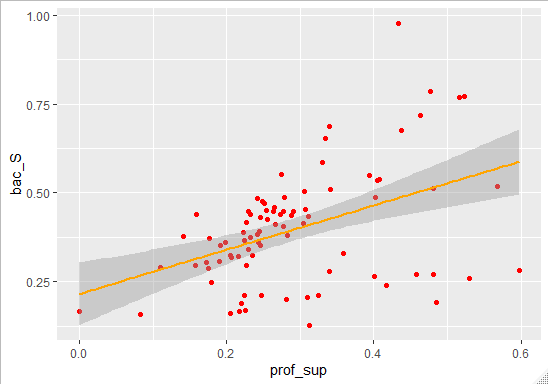
Les données utilisées pour étudier l’évolution de la composition sociale des populations étudiantes au sein des universités françaises sont issues des bases du SISE et du SIES entre 2007 et 2015. Le système d’information SISE recense, annuellement depuis 1995, les étudiants inscrits dans une centaine d’établissements publics (universités et instituts) supérieurs en France métropolitaine et outre-mer. S’agissant d’inscriptions administratives, la présence effective de l’étudiant n’est pas indiquée. Cela peut biaiser les résultats puisque des étudiants pourraient s’inscrire pour toucher une bourse par exemple et dans ce cas privilégieraient une université pour sa localisation ou sa sélection à l’entrée. De cette base de données sont tirées les informations suivantes : établissement dans lequel l’étudiant est inscrit, catégorie socio-professionnelle du parent référent, sexe de l’étudiant, diplôme de série du bac, niveau d’étude actuel (s’il est inscrit en licence, master ou doctorat) et académie à laquelle l’établissement est rattaché. De ses informations sont déduites les effectifs d’étudiants au sein des établissements étudiés ainsi que les taux relatifs aux genre, catégories socio-professionnelles du parent référent, série de bac et niveau de formation. Il est dommage que les informations relatives à l’attribution d’une bourse et de son échelon ne soient pas disponibles sur la période étudiée puisqu’il s’agirait d’en apprendre plus sur le niveau d’aisance de l’étudiant, la bourse étant basée sur des critères sociaux tels que le montant de la rémunération des parents ou le nombre de frères et sœurs en études supérieures, ce qui semble être un indicateur pertinent quant à l’origine sociale d’un étudiant. Les catégories socio-professionnelles ont été découpées en 7 catégories, ce qui exclue d’une part les détails des métiers et forme des ensemble plus hétérogène, et d’autre part certaines catégories qui n’ont pas pu être prises en compte. Sachant que la catégorie socio-professionnelle d’un parent est plus indicative que représentative du milieu social auquel l’étudiant appartient, de l’éducation qu’il a pu recevoir ou encore des ressources auxquelles il peut accéder ; ces données permettent de se rapprocher de ce que peut être son origine sociale mais n’apporte pas une information précise pour toutes ses raisons.

A partir de cette base de données, les recherches effectuées ont mises en évidence la présence d’une polarisation entre les établissements accueillant plus d’étudiants d’origine favorisée et celles composées d’étudiants d’origine plus populaire. Et également son accroissement entre 2007 et 2015.

Pour ce qui est du classement ou non, et du rang des universités Françaises, ces données sont tirées de l’ARWU (Academic Ranking of World Universities).

# Graphique du taux de bacheliers S en fonction du taux d’étudiant dont le parent référent exerce une profession supérieure

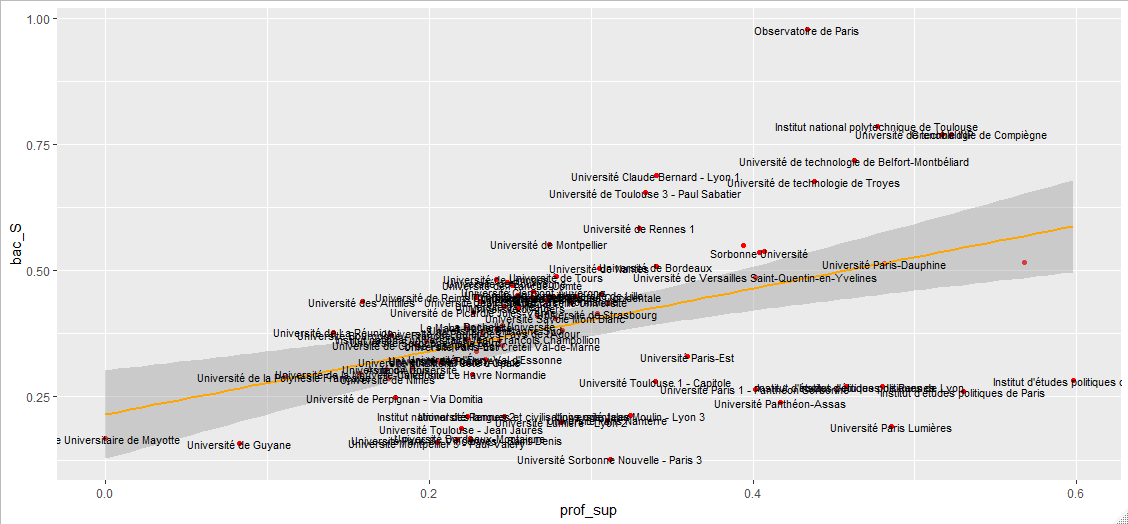
Figure 1: taux de bacheliers S en fonction du taux de parents référents exerçant une profession supérieure



Sur ce premier graphique, nous pouvons observer une corrélation positive entre les variables bac\_S et professions supérieures, ce qui signifie que ces deux variables tendent à augmenter ensemble. Le coefficient de corrélation calculé sur RStudio arrondi au millième est 0.437, compris entre 0 et 0,5 et proche de 0,5, elle est modérément positive.

Une université comprenant dans ses effectifs une part d’étudiants donc le parent référent exerce une profession supérieure importante aurait donc probablement une part plus importante de diplômés du bac série S. Nous pouvons voir tout de même un fort taux de dispersion qui s’accentue à mesure que les taux de professions supérieures augmentent. En rajoutant le libellé de l’établissement, ce qui rend le graphique peu lisible, nous pourrons mettre en lien les variables bac S et professions supérieures avec les composantes des universités.

Figure 2: taux de bac S en fonction du taux d'étudiants dont le parent référant exerce une profession supérieure



Sur la gauche du graphique, on observe les universités de DOM-TOM avec un très faible taux d’étudiants dont le parent référent exerce une profession supérieure et un très faible taux de bac S. A l’extrême droite du graphique, pour les établissements à fort taux de bac S , ce sont les établissements à composante scientifique qui sont représentés, on y voit les universités technologiques ainsi que les instituts polytechniques. Ces établissements comprennent un taux également plus élevé d’étudiants dont le parent référent exerce une profession supérieure.

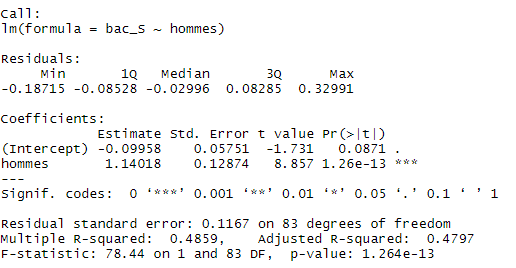
Comme la dispersion augmente avec le taux de professions supérieurs, nous retrouvons également à l’extrême droite du graphique des universités à composante littéraire, majoritairement situés à Paris et renommées et celles qui ont les plus forts taux de professions supérieures sont les instituts d’études politiques.

# Modèle linéaire expliquant le taux de bacheliers S par la part des hommes dans la population étudiante et les différentes origines sociales.

Dans un premier temps, nous verrons les analyses bivariées du taux de bachelier S avec la part d’hommes, puis du taux de bachelier S avec 3 différentes origines sociales ; puis nous réaliserons le modèle linéaire avec les variables explicatives du taux de bachelier S dans les établissements d’étude supérieures.

## Analyses bivariées

Figure 3: Sortie 1 de Rstudio

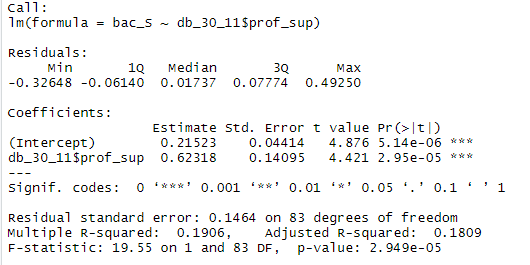


Le taux d’hommes est une variable très explicative du taux de bacheliers S dans une université, en effet la t-value est très faible ( 1,26e-13 donc proche de zéro. Le coefficient de variation est positif donc dans les universités à fort taux de bac S (plutôt scientifiques du coup) on trouve également un plus fort taux d’hommes. On remarque aussi que le coefficient de détermination est proche de 0,5 (R^2 = 0,4859), le modèle choisit (taux d’hommes) détermine pas loin de 50% des points (taux de bac S).

Pour ce qui des différentes origines sociales, je vais faire des analyses bivariées avec 3 groupes de métiers que j’ai sélectionnés de catégories sociales assez différentes : les professions supérieures, les employés et les ouvriers.

### Profession supérieure

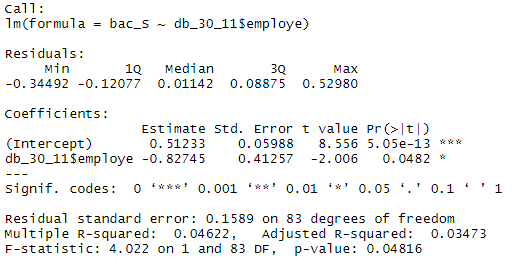
Figure 4: Sortie 2 de Rstudio



La corrélation entre e taux d’étudiants dont le parent référent exerce une profession supérieure et le taux de bacheliers S dans une université est statistiquement très significative. En effet la t-value est très faible (2,949 e-05), et les résidus étant compris entre -2 et 2 , il n’y a pas d’écart significatif. Nous avons également vu plus haut que la corrélation entre ces deux variables est positive et nous voyons ici que l’écart type du coefficient le fait rester dans les positifs.

### Employé

Figure 5: Sortie 3 de RStudio

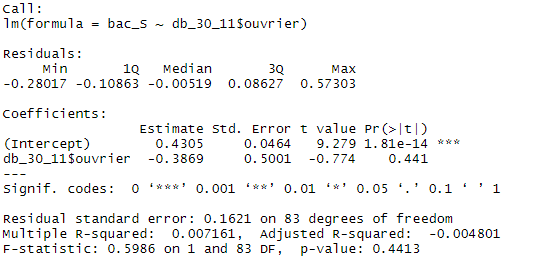


Le taux d’étudiants dont le parent référent est employé est une variable explicative du taux de bacheliers S dans une université, en effet la p-value est de 0,0482 donc tout juste en-dessous de 0,05 le seuil de risque α d’accepter la variable alors qu’elle serait fausse.

Les résidus quant à eux sont compris entre -2 et 2, il n’y a pas d’écart significatif. Le coefficient de corrélation calculé est négatif (-0.215 arrondi au millième) et ce même avec la prise en compte de l’écart type, donc les variables bac S et employé tendent à diminuer ensembles.

### Ouvrier

Figure 6:Sortie 4 de RStudio



Le taux d’étudiants dont le parent référent est ouvrier est une variable non explicative du taux de bacheliers S dans une université, en effet la p-value est de 0,441 donc bien supérieure au seuil à partir duquel nous rejetons une variable soit 0,05.

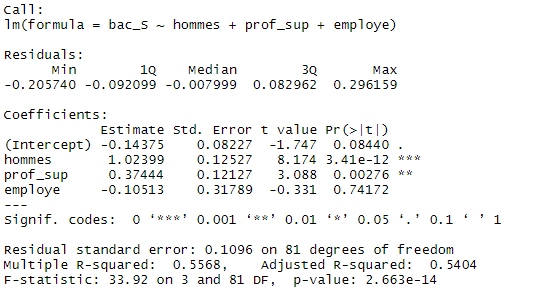
Les résidus quant à eux sont compris entre -2 et 2, il n’y a pas d’écart significatif. Nous ne pouvons pas déterminer si le coefficient de corrélation est positif ou négatif car il est compris entre [-0,887 ; 0,113].

Ces analyses bivariés mettent en avant le fait qu’un étudiant aisé (dont un des parents exerce une profession supérieure) augmente ses probabilités de détenir un bac S dans un établissement d’études supérieures, puisque nos données proviennent des inscriptions en étude supérieures. Ce n’est pas le cas pour des étudiants d’origine plus populaires, en effet les étudiants dont le parent référent est employé diminue sa probabilité de détenir un bac S ; tandis que pour les enfants d’ouvrier, il n’y a pas de corrélation linéaire ni même de corrélation significative ce qui peut s’interpréter comme un manque de ce type de donnée pour calculer une éventuelle corrélation entre ces variables soit peu d’étudiants dont le parent est ouvrier, qui aurait validé un bac S et qui serait inscrit dans un établissement d’études supérieures.

## Analyse multivariée et Modèle linéaire

Pour réaliser le modèle linéaire du taux de bachelier S, nous utiliserons la variable du taux d’hommes ainsi que les variables professions supérieures et employé puisqu’une première analyse bivariée à validé l’explicabilité de ses origines sociales quand au taux de bachelier S au sein des établissements.

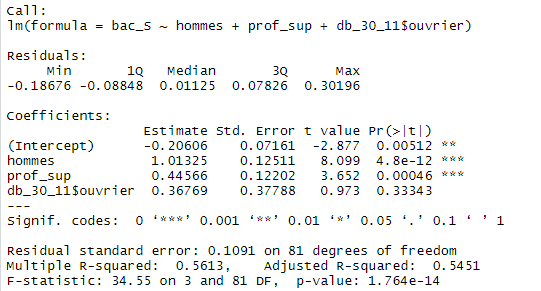
Figure 7: Sortie 5 de RStudio



Le modèle multivarié valide la corrélation positive et significative entre les variables hommes et professions supérieures et le taux de bac S, la variable employée quant à elle n’est plus explicative. Le R^2 = 0,557 donc notre modèle explique presque 56% de la variation de la variable bac S.

Nous testons une autre variable pour voir ce que cela aura comme effet et si cela modifie les variables importantes.

Figure 8: Sortie 6 de RStudio



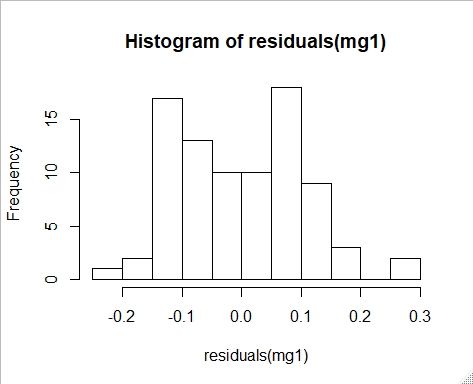
Lorsque l’on remplace la variable employée par la variable ouvrier cela ne fait que renforcer la significativité de la corrélation entre les variables professions supérieures et bac S.

Nous reprenons le premier modèle élaboré, avec les variables hommes, professions supérieures et employé. Le modèle suppose que la régression est linéaire, les termes d’erreurs ont même variance, ils sont indépendants et issus d’une loi normale. Avant d’interpréter les résultats il nous faut donc vérifier ces hypothèses. Nous testons la linéarité du modèle avec un test de Rainbow. La p-value de mg1 est de 0,8541 > 0,05 donc le test de Rainbow valide la linéarité des modèles. Nous vérifions ensuite que le modèle est globalement significatif. La p-value associée au f-stat (=2.663e-14) de notre modèle est inférieure à 0,05 on en déduit que notre modèle est globalement significatif. Il faut maintenant tester les hypothèses sur les résidus, à savoir : L’homoscédasticité des résidus, la normalité des résidus et l’absence d’autocorrélation des résidus.

Nous allons vérifier à l’aide d’un test de breush-pagan l’hypothèse d’homoscédasticité (V(ui)= 6^2) des résidus. BP = 14.091 > 0,05 donc nous rejetons l'hypothèse d'hétéroscédasticité et supposons l'homoscédasticité des résidus.

Pour vérifier la normalité des résidus, nous commençons par une analyse graphique.

Figure 9:Fréquence des résidus du Modèle 1



La distribution ne ressemble pas à une distribution normale des résidus. Et en effet le test de Shapiro ne valide pas la normalité des résidus (la p-value est inférieure à 0,05, on rejette l’hypothèse de normalité des résidus),c’est parce qu’il y a des différences significatives de variance entre les groupes.

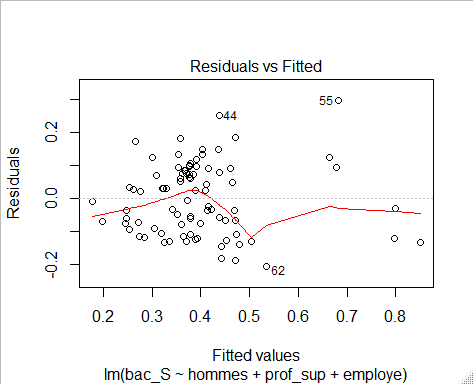
Enfin le test de colinéarité VIF nous indique que notre modèle général ne comporte pas de variables auto-corrélées tout comme les modèles auxiliaires (hommes~prof\_sup+employe, prof\_sup~employe+hommes et employe~hommes+prof\_sup) ne comportent pas d’autocorrélation. Le résultat du test est proche de 1 donc les facteurs ne sont pas influencés par la corrélation avec d'autres facteurs.

Tableau 1 : Résultat des tests

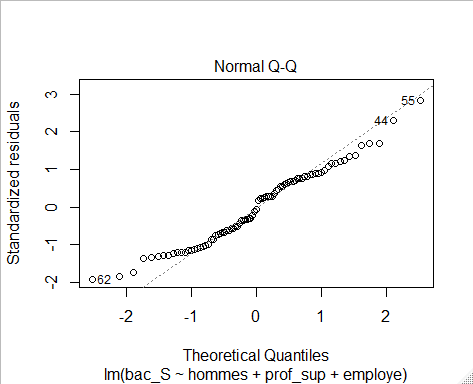
|  |  |
| --- | --- |
| **Test d’hétéroscédasticité**  **(Breusch Pagan)** | Accepte H0 |
| **Test de linéarité (Rainbow)** | Accepte H0 |
| **Test de Fisher** | Modèle globalement significatif |
| **Test de normaltié des résidus (Shapiro)** | Rejette H0, non normalité des résidus |
| **Test de VIF** | Pas d’autocorrélation |

Toutes les hypothèses sont validées sauf l’hypothèse de normalité des résidus, nous pouvons garder notre modèle mais le risque d'erreur de type 1 (rejeter l’hypothèse nulle à tort) peut être plus important ou plus faible que la valeur escomptée α = 5%.

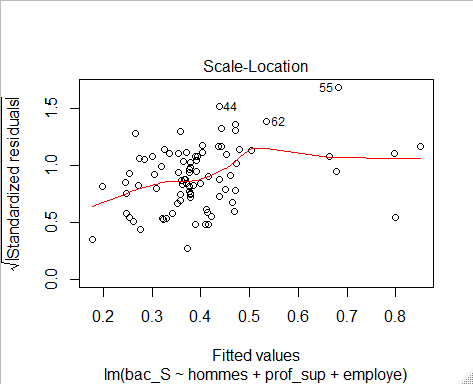
### Analyse graphique des estimations du modèle



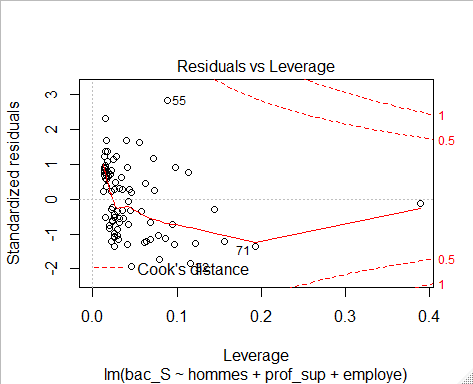
Ici, nous voyons que la linéarité semble tenir raisonnablement bien malgré des chutes ou montées de la ligne rouge assez éloignées de la ligne pointillée. Les points 45, 55 et 62 peuvent être des valeurs aberrantes, avec des valeurs résiduelles importantes.



Nous observons ici que les points sont surtout au niveau de la zone centrale, nous avons une distribution normale des valeurs. Ils ne suivent pas vraiment la ligne centrale donc ne suivent pas une loi normale, c’est ce que nous avons trouvé avec le Shapiro test. Les valeurs 55 et 62 sont de nouveau assez éloignées des autres auquel on rajoute l’élément 44.



Les résidus ne suivent pas totalement les valeurs attendues, on peut voir notamment qu’une large part des établissements qui sont représentés par les points à gauche contribuent moins à expliquer la variable contrairement à celles de droite qui forment une ligne proche de 1. Les mêmes extrêmes apparaissent : 44, 62 et 55.



Ce graphique nous permet de repérer les individus extrêmes donc le 55 et le 62 de nouveau.

Comme nos points sont très dispersés, ces graphiques ne nous apportent que peu d’informations, il est difficile de juger les extrêmes aux vues des fortes disparités.

Aux vues de ces résultats, un établissement composé d’une majorité d’hommes et d’étudiants d’origine sociale aisée -si nous considérons qu’un parent exercent une profession supérieure garantit une origine sociale aisée- comporterai un taux d’étudiants détenant un bac S plus élevé.

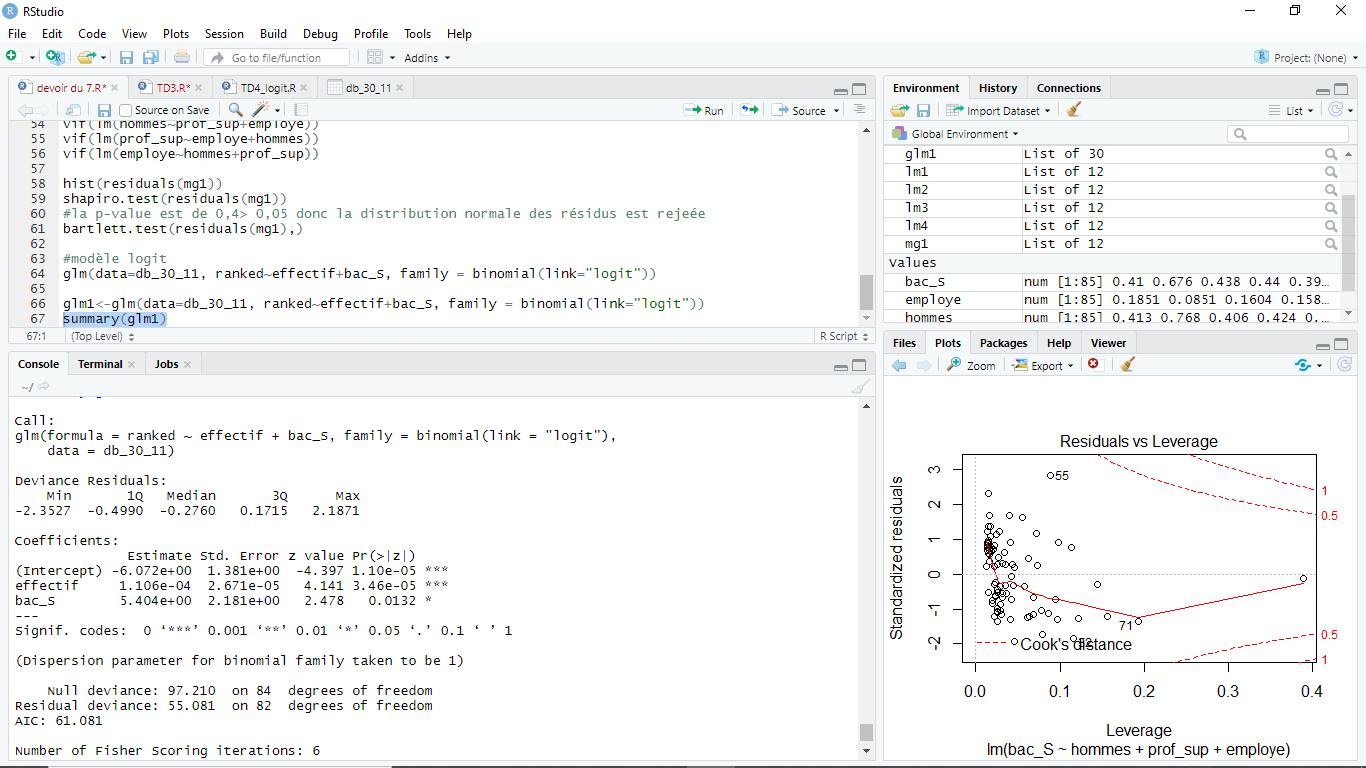
# Faire un modèle logit expliquant le fait d’être classé à Shanghai à l’aide des variables explicatives de votre choix.

On peut supposer que le taux de bachelier explique en partie le fait d’être classé à Shangaï puisque le classement de Shangaï est élaboré en prenant en compte la publication d’articles et les prix remportés ce qui peut favoriser les instituts polytechniques et les universités à composantes scientifiques qui sont mécaniquement composées de fort taux de bacheliers S.

La variable « effectif » est quant à elle sélectionnée pour voir si elle influence sur la probabilité d’être classé puisque le gouvernement Français pousse à la fusion d’universités et d’écoles pour créer des grandes universités qui boosteraient la recherche ce qui leur permettraient de remonter dans le classement[[1]](#footnote-1).

Pour effectuer ce modèle logit, nous allons donc prendre les variables « effectif » et « bac S » pour vérifier nos hypothèses précédemment énoncées.

Figure 10: Sortie 7 de RStudio



La p-value de l’effectif est proche de zéro donc la corrélation est significative, la p-value des bac S est proche de 0,01 donc inférieure au seuil d’erreur de 0,05 donc cette variable est également significative.

Les signes des coefficients sont tous les deux positifs donc plus l’effectif est grand plus on a de chances d’être classé, de même avec le taux de bas S, s’il est élevé on a plus de chance d’être classé. Le coefficient de l’effectif peut sembler très faible mais c’est parce que les effectifs sont en milliers.

Enfin, nous allons visualiser ces résultats à l’aide de graphiques.

Figure 11: Universités classées ou non à Shangaï en fonction de l’effectif et du taux de bac S

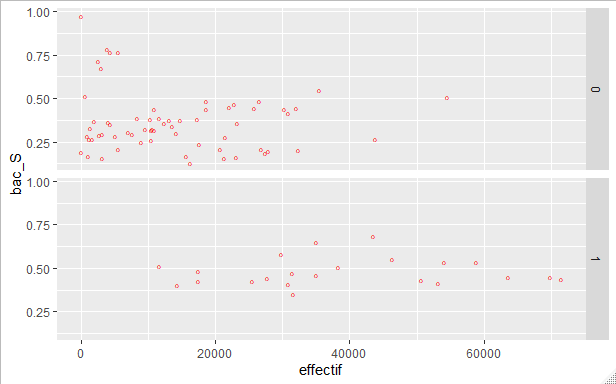
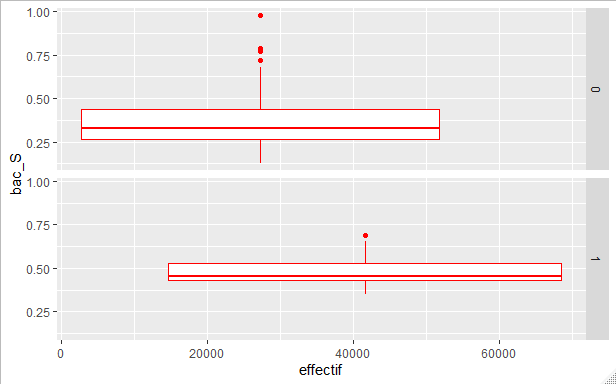


Figure 12: Boite à moustache des universités classées et non classées à Shangaï en fonction de l’effectif et du taux de bac S



Avec 0 =pas classé et 1 = classé

Nous pouvons voir que la répartition des effectifs des établissements classés est assez dispersée mais ne descends pas en dessous du seuil des 10000 étudiants tandis que les établissements non classés ont au contraire des effectifs majoritairement en dessous des 35000. Pour ce qui est du taux de bac\_S, on peut voir sur le deuxième graphique que la médiane des taux de bac\_s pour les établissements non classés est environ de 35% tandis que pour les établissements classés elle est proche des 50%, sur le premier graphique on peut même voir que ce taux ne descend pas en dessous des 30% dans les établissements classés. On observe quelques valeurs extrêmes dans les établissements non classés, il s’agit d’établissements à très fort taux de bac S mais dont les effectifs sont très faibles (<10000).

Ces graphiques corroborent les résultats que nous avons trouvé pour le modèle logit.

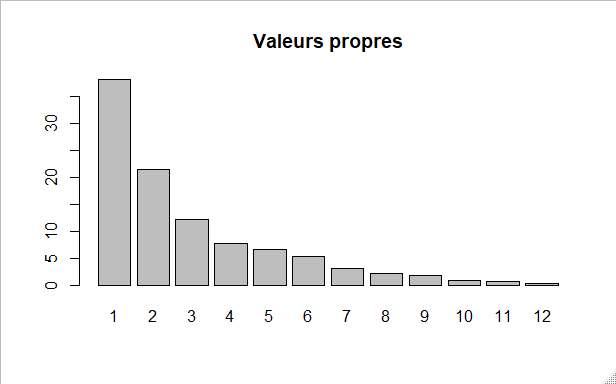
Nous pouvons en plus rappeler que dans la partie précédente, nous avons vu que le taux de bac S était corrélé positivement aux étudiants masculins et d’origine sociale aisée. Les établissements scientifiques et à fort effectif sont donc plus susceptible d’être classé à Shangaï ce qui induit également des établissements dans lesquels il y aurait plus d’hommes et d’individus dont le parent référé a moment de l’inscription exerce une profession supérieure plutôt qu’ouvrier ou employé.

# Analyse en Composante principale

Nous utiliserons la méthode d’analyse en composante principale pour transformer nos variables corrélées en composantes principales qui nous permettrons de représenter nos données, les universités, dans un espace construit en deux axes indépendants expliquant au mieux leur variabilité.

Pour effectuer notre PCA, nous sélectionnons les variables qui nous semblent les plus pertinentes pour mettre en avant l’inégale répartition de la composition sociale selon les universités françaises. La première figure porte sur l’éboulis des valeurs propres, elle se présente sous forme de graphique en barre et nous donne le pourcentage d’informations contenu dans chaque dimension. Nous utiliserons donc les variables précédemment utilisées portant sur le métier des parents [chef\_ent, prof\_sup, art\_comm, agriculteur, employe, ouvrier, prof\_inter] et le type de bac passé [bac\_S, bac\_L, bac\_ES, bac\_pro et bac\_techno]. Les valeurs des variables sont pondérées par les effectifs car certaines universités ont un effectif beaucoup plus petit et pèsent aussi lourd que les très grandes donc en pondérant nous faisons en sorte que les petites influent moins sur la construction des axes.

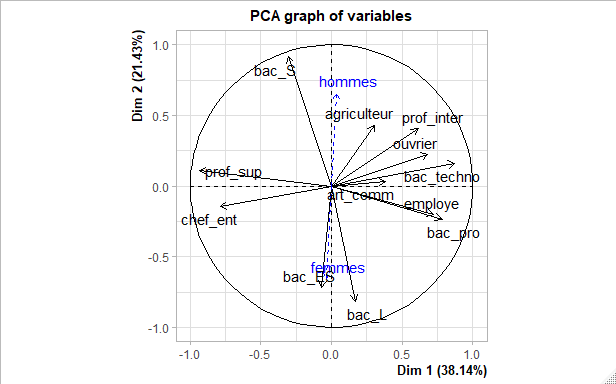
Figure 13: éboulis des valeurs propres



Pour choisir le nombre d’axes factoriels et vérifier leur pertinence, nous nous baserons sur deux critères. Le premier est le critère du coude qui consiste à sélectionner les axes se trouvant avant un décrochement en forme de « coude ». Nous n’avons pas clairement ce type de décrochement sur notre éboulis puisque nous observons un première dimension très explicatives suivie d’une décroissance plus proche d’une courbe que d’un coude. Nous pouvons nous appuyer sur le critère de Kaiser pour retenir les axes dont l’inertie est supérieure à l’inertie moyenne à savoir supérieure à 8,33%. Les premières dimensions répondent bien à ce critère, la troisième est légèrement au-dessus. Pour faciliter les analyses à venir, nous ne retiendrons que les deux premiers axes pour obtenir le premier plan factoriel car il compréhensible à l’œil comme nous verrons dans les figures suivantes. De plus il ne déforme pas trop le nuage car il explique près de 60% (exactement 59,57%) de l’inertie totale et est pertinent à interpréter puisqu’il synthétise une bonne partie de l'information.

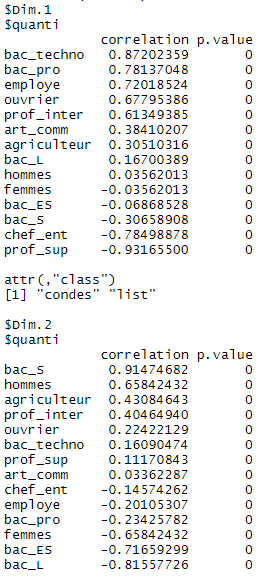
Le graphique de la PCA reprend les mêmes variables que précédemment pondérées par l’effectif, auxquelles on ajoute les variables qualitatives supplémentaires liées au genre [hommes, femmes] qui ne seront donc pas utilisées pour sa construction.

Figure 14 : graphique de la PCA des variables



Axe 1

On observe que la dimension 1 regroupe à gauche les variables professions supérieures et chef d’entreprise et à droite les ouvriers, employés , professions intermédiaires et artisans commerçants, on peut donc interpréter l’axe 1 comme un axe d’origines sociales. On peut vérifier nos suppositions en regardant quelles variables servent le plus à la construction des axes :



On remarque bien que les origines plus populaires (ouvriers, employés et professions intermédiaires) sont les plus corrélés positivement tout comme les origines les plus aisées (chef d’entreprise et profession supérieure) sont les plus corrélées négativement à l’axe horizontal. A la lecture de ce graphique, nous pouvons nous attendre à ce que les universités concentrant de plus forts taux d’étudiants d’origine populaires seront représentées à droite sur l’axe 1. Nous verrons la projection des universités en figure 15.

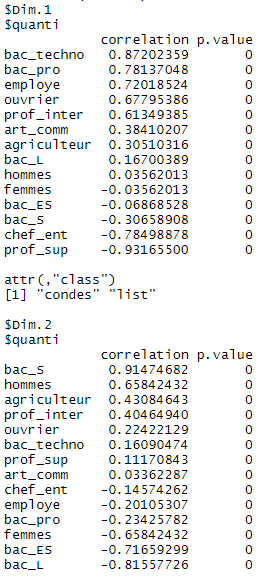
Contrairement aux bacs S, L et ES qui vont contribuer à l’axe 2 et dont on parlera après, on observe que les bacs professionnels et technologiques sont fortement corrélés positivement à l’axe 1, c’est-à-dire aux origines sociales d’après notre interprétation de la construction de l’axe en fonction des catégories socio-professionnelles du parent référent. Ainsi les universités ayant un pourcentage plus élevé de ce type de bac dans leurs rangs seront situées très à droite, comme pour celles qui ont une proportion plus élevée d’ enfant d’employés et d’ouvriers tandis que les universités à majorité de bac S sont plus à gauche comme pour les professions supérieures et les chefs d’entreprise.

On sait en effet que les grands lycées généraux (situés plutôt à Paris et dans les grandes villes) ont des préférences pour le bac S tandis que les lycées de banlieue vont proposer plus souvent des bac pro et techno.

La faible corrélation des métiers indépendants (artisan/commerçant et agriculteur) avec l’axe 1 peut s’interpréter comme une sous-représentation de ses métiers dans les catégories socio-professionnelles des parents d’étudiants ce qui ne permettrait pas en termes d’analyse d’établir un lien entre ces catégories socio-professionnelles et la composition sociale des universités.

Axe 2

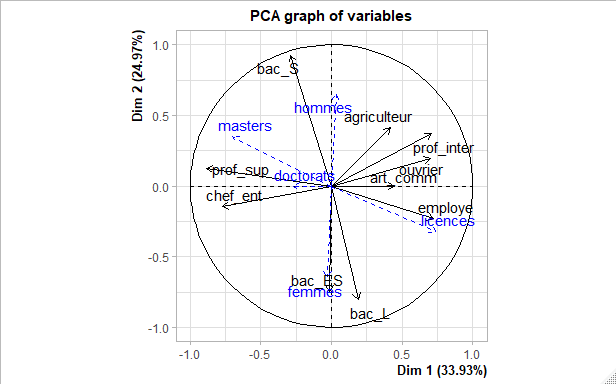
Comme supposé lors du choix des variables supplémentaires, hommes et femmes n’influent pas sur l’origine sociale puisque les flèches sont plutôt à la verticale (donc sur l’axe 2) mais ce résultat témoigne d’une influence du genre de l’étudiant dans son orientation d’études. L’axe 2 semble se construire à partir des types de bacs passés, on aura en haut des universités a composantes scientifiques et en bas des universités plutôt littéraires et commerciales, ces dernières étant moins corrélées négativement à l’axe 2 que les bacs L. Nous pouvons vérifier nos hypothèses avec les corrélations positives et négatives fortes qu’ont respectivement les bacs S et L avec l’axe 2.



L'axe 1 résume plus de 38 % de l'information du nuage initial ce qui signifie que l’origine sociale des étudiants apparait comme une mesure pertinente de la composition sociale des universités françaises et qu’une opposition entre les origines aisées et populaires apparait effectivement.

Nous allons rajouter le cursus des études [licence, master et doctorat] en variables supplémentaires afin de voir si les origines sociales aisées favorisent la poursuite d’études en supérieures. Le résultat graphique que nous obtenons est le suivant :

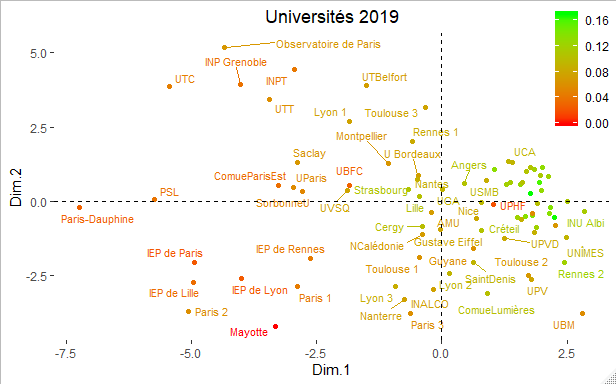
Figure 15:Graphique de PCA avec les cursus



Nous observons que la variable licence est corrélée positivement à l’axe 1 et négativement au deuxième tandis qu’à l’opposé, la variable master l’est négativement au premier et positivement à l’axe 2. La variable doctorat est quant à elle faiblement corrélée aux axes. Cela nous permet de définir un rapport entre la poursuite en études supérieures et l’origine sociale de l’étudiant, les universités composées d’étudiants plus aisées seront celles qui présenteront des taux d’étudiants en Master plus élevés, à l’inverse, les universités situées à gauche de l’axe 1, composées de taux plus élevés d’étudiants d’origines populaires seront celles qui présentent des taux d’étudiants en Licence plus élevés. Le type d’études poursuivie (plutôt scientifiques ou littéraires) influent également sur la poursuite ou non en études supérieures avec une poursuite en Master corrélée positivement avec l’axe 2 et donc que l’on retrouvera plus pour les études scientifiques.

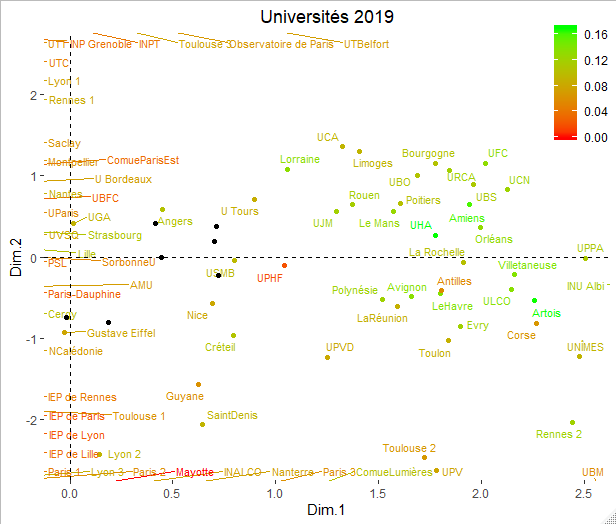
Nous allons maintenant vérifier nos interprétations en faisant apparaitre les universités sur un graphique qui reprend les mêmes axes. Nous mettons en couleur une des variables pour mettre en avant la composition sociale des universités que l’on pourra voir ici en fonction du taux d’étudiants dont le parent référent est ouvrier.

Figure 16 : représentation des universités dans le plan



J’effectue un zoom sur les universités agglutinées en vert à gauche pour rendre leur libellé lisible :

Figure 17: zoom sur le plan



Selon nos hypothèses précédentes, les universités des plus aux moins composées d’étudiants d’origines aisées seront situées depuis la gauche du graphique jusqu’à la droite. Cette hypothèse s’observe bien avec l’échelle de couleur, à gauche les universités ne comportent que très peu d’ouvriers et plus on va vers la droite plus on en trouve. On observe également que les grandes universités parisiennes et les instituts d’études politiques sont plus à gauche sur l’axe 1 donc composés d’origines sociales plus aisées ; au milieu, on trouve des universités de Province ou de Paris moins réputées et dont la composition est plus homogène ; enfin tout à droite se trouvent des universités situées en banlieue et des universités de province et de Paris plus axées sur les lettres et l’art qui sont composées d’étudiants ayant des origines sociales plus populaires.

Nous observons bien que les universités à dominante littéraire apparaissent en bas du graphique tandis que les scientifiques sont les plus en haut. Et cela fait également ressortir le fait que les universités les plus populaires sont situées plus en bas, c’est-à-dire qu’elles sont composées d’étudiants ayant passé plutôt des bacs littéraires et commerciaux. Cela peut s’interpréter comme le fait que les étudiants d’origine sociale moins aisées vont se tourner vers des études moins scientifiques et donc impacter la composition sociale des universités selon leurs composantes.

1. Nunès, Eric. « Le classement de Shanghaï n’est pas fait pour mesurer la qualité des universités françaises », Le Monde, 15/08/2019. Disponible [en ligne](https://www.lemonde.fr/campus/article/2019/08/15/le-classement-de-shanghai-n-est-pas-fait-pour-mesurer-la-qualite-des-universites-francaises_5499548_4401467.html) [↑](#footnote-ref-1)