Projet séminaire d'intégration

Axel CHABE / Kéroudine BELLADJO

22/11/2022

1. Construction de la base à étudier.

Commençons par charger les librairies qui nous seront utiles ainsi que notre base de données ODD_DEP et le Tableau_propre que nous vous avons fourni.

```
library("FactoMineR")
library("factoextra")
library("missMDA")
library("corrplot")
library("car")

ODD_DEP <- read.csv( file = "ODD_DEP.csv", sep = ";")
Tableau_propre <- read.csv( file = "Tableau_propre.csv", sep = ";")
```

Tout d'abord, nous avons trouvé que les données de l'INSEE étaient très mal présentées. C'est pourquoi nous avons décidé de réorganiser toutes leurs données de l'année 2018. Pour ce faire, nous avons créé un tableau avec pour lignes chaque département et pour colonnes toutes les combinaisons différentes variables/sous-champs.

```
Tab_2018 <- ODD_DEP[complete.cases(ODD_DEP$A2018),]
#Retire les lignes où il n'y a pas de valeur en 2018 pour 'accelerer' la futur boucle.

mat_champs_sschamps <- matrix(c(Tab_2018$variable[1:nrow(Tab_2018)],
                                Tab_2018$sous_champ[1:nrow(Tab_2018)]),ncol = 2)
uni_mat_champs_sschamps <- unique(mat_champs_sschamps)
#On a toutes les combinaisons variables/sous-champs uniques.

Tableau_propre <- matrix(NA,nrow = 102,ncol = nrow(uni_mat_champs_sschamps))
rownames(Tableau_propre) <- c(" ",unique(ODD_DEP[,2][1:706]))
colnames(Tableau_propre) <- uni_mat_champs_sschamps[,1]
Tableau_propre[1,] <- uni_mat_champs_sschamps[,2]
```

Ensuite, nous remplissons ce tableau vide avec les valeurs de ODD_DEP.
Dans les deux versions précédentes de notre rendu, il était impossible de faire tourner cette boucle en une seule fois et cela durait plusieurs heures.
Pour le rendu final, nous avons essayé de l'optimiser même si cela n'est pas encore totalement satisfaisant à nos yeux. En effet, même si maintenant nous pouvons faire tourner la boucle en une seule fois, cela dure encore une quinzaine de minutes dans ce markdown (5 minutes dans un simple fichier vide). C'est pourquoi nous vous avons demandé de charger directement le 'Tableau_propre' et de ne pas lancer la boucle.

On enregistre maintenant ce tableau contenant toutes les données de l'Insee d'ODD_DEP de l'année 2018 triées afin de ne pas avoir à relancer la boucle. Une fois le tableau créé, il suffit juste de le recharger.

```
write.table(Tableau_propre,"Tableau_propre.csv",sep=";")
Tableau_propre <- read.csv( file = "Tableau_propre.csv", sep = ";")
```

Ensuite, nous avons sélectionné 20 variables que nous avons trouvé intéressantes parmis les 600 disponibles.

2. Analyse descriptive de nos données.

a) Imputation des données:

Notre tableau maintenant constitué des 20 variables, nous voulons regarder la corrélation entre les variables avec `cor(Tableau_final)` afin de voir quelles variables nous allons garder. Cependant, il manque les données de certaines variables pour certains DOM. La solution la plus simple aurait été de les supprimer et de ne pas les analyser du tout mais nous n'avons pas voulu mettre de côté ces départements. Il nous semblait en effet important de parler un minimum de ces départements trop souvent négligés.

Nous avons donc décidé d'imputer les valeurs manquantes.
Nous aurions aussi pu utiliser la moyenne des valeurs présentes mais cela n'aurait pas été très correct car la moyenne est trop sensible aux comportements extrêmes par exemple.
Pour réaliser cette imputation, nous avons utilisé le package 'missMDA' qui va nous permettre d'imputer les valeurs manquantes de manière plus cohérente.

Cependant, nous analyserons les DOM séparément dans un second temps, car le comportement associé aux variables n'est en rien comparable avec celui des départements de la France métropolitaine.

```
nb <- estim_ncpPCA(Tableau_final,ncp.min=0, ncp.max=5, method.cv="Kfold")
```

```
##   |                                                                      |
```

```
res_impute <- imputePCA(Tableau_final,ncp=nb$ncp)
```

Pour ce faire, nous avons besoin d'estimer le nombre de dimensions qui sera utilisé pour compléter le tableau de données.
Nous avons donc utilisé la commande`nb <- estim_ncpPCA(Tableau_final, method.cv="Kfold")` voulant dire qu'on peut tester le nombre de dimensions entre 0 et ncp.max = 5. (Plus ne serait pas forcément utile et long à faire tourner.)
Cette procédure est très longue car pour chaque cellule la valeur est enlevée, estimée avec zéro composante, une composante, deux composantes ... jusqu'à cinq composantes. Et on calcule l'écart entre la valeur imputée avec zéro composante et la valeur réelle, pour chaque nombre de composantes et pour chaque cellule. A chaque fois l'algorithme d'imputation est utilisé ce qui est très long.
Le nombre de composantes estimé optimale est donné avec la commande `nb$ncp` donc c'est ce nombre que nous allons utiliser pour imputer les données dans le jeu de donnée.
Notre tableau de données étant assez conséquent, nous avons utlisé la méthode "Kfold"

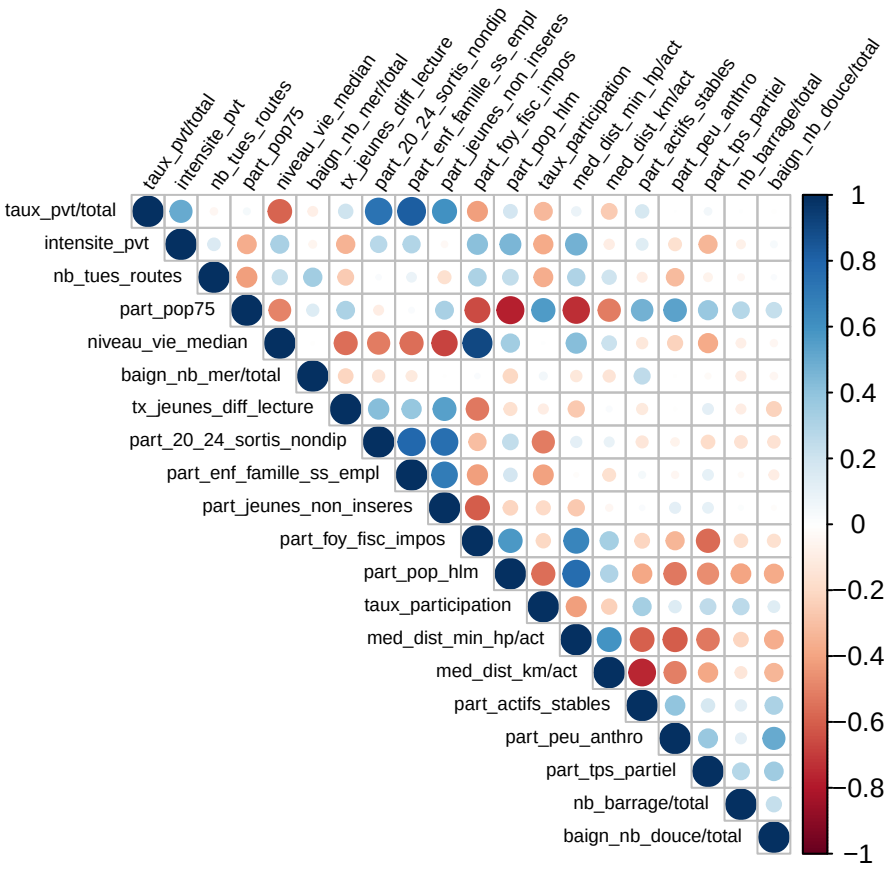
b) Choix des variables et définitions:

Pour visualiser notre tableau, il nous suffit d'utiliser `res_impute$CompleteObs`.

Regardons la corrélation entre nos variables en ne choisissant que les individus de la France métropolitaine.

```
mat_cor <- cor(res_impute$completeObs[1:96,])  #Pas très visuel car beaucoup de variables.
```

```
corrplot(mat_cor, type="upper", order="original", tl.col="black", tl.srt=55, tl.cex=0.6)
```



Remarque: Pour faire la matrice de corrélation, nous enlevons les DOM encore une fois car leurs valeurs (assez extrêmes) perturbent les corrélations de certaines variables.

En visualisant de cette manière notre matrice de corrélation, nous allons pouvoir choisir beaucoup plus facilement les variables à garder et celles à enlever.
Par exemple, il parait évident de ne pas garder la variable 'baign_nb_mer_total' ou encore la variable 'nb_tues_routes' qui sont très mal corrélées avec les autres.

Au final, nous gardons les variables: - Taux pauvreté total, (taux_pvt/total)
- Part de la population ayant au moins 75ans (part_pop75)
- Médiane du niveau de vie (niveau_vie_median)
- part des 20-24ans non diplômés (part_20_24_sortis_nondip)
- Part des 0-17ans dans une famille sans actif occupé (part_enf_famille_ss_empl) - Part des foyers fiscaux imposés (part_foy_fisc_impos) - Part des jeunes (18-25ans) non insérés (part_jeunes_non_inseres)
- part des foyers fiscaux imposés (part_enf_famille_ss_empl)
- Part de la population dans les HLM (part_pop_hlm)
- Participation au premier tour des élections législatives (taux_participation)
- Distance médiane des navettes domicile-travail pour les actifs (med_dist_min_hp/act)
- Part des actifs stables (part_actifs_stables)

Remarque: Ce choix est quelque peu arbitraire mais nous avons choisi les variables plus intéressantes à nos yeux.

Pour cela nous avons utilisé la commande:

```
res11var <- res_impute$completeObs[,c(-2,-3,-6,-7,-15,-17,-18,-19,-20)]
```

c) Analyse descriptive:

Nous allons maintenant expliquer brièvement les informations que nous donne chaque variable sur les régions ainsi que leur pertinence. Pour cela nous nous sommes aidés des fichiers d'aide à la compréhension des indicateurs disponible sur https://www.insee.fr/fr/statistiques/4505239#documentation en bas de page.

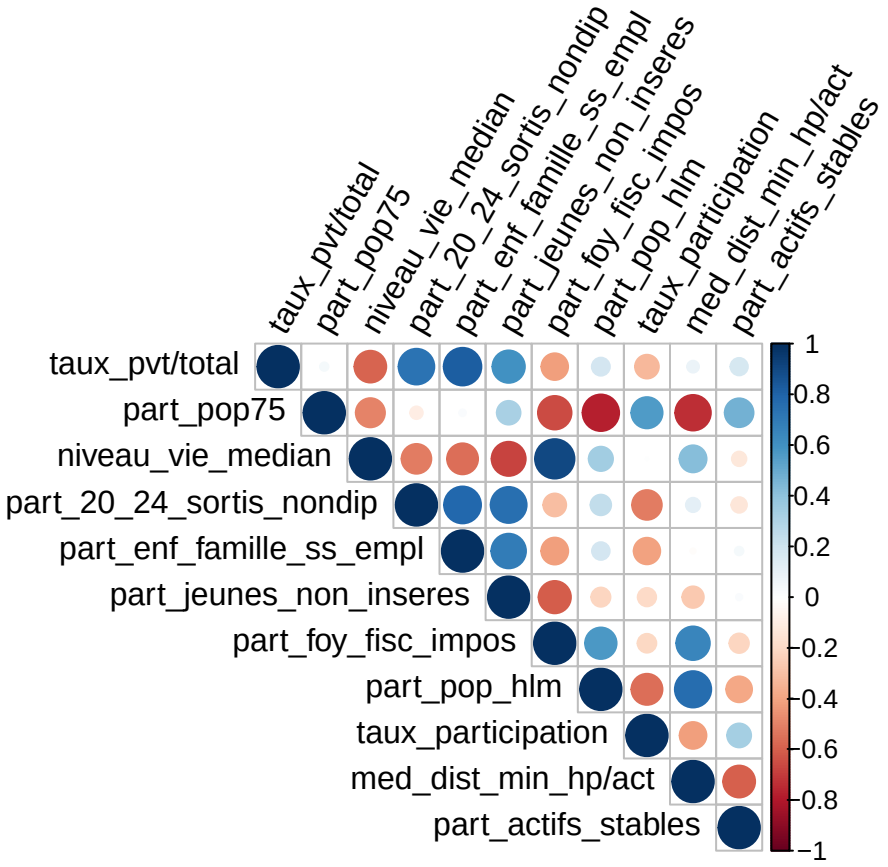
- taux_pvt/total (€): Le taux de pauvreté est le pourcentage de la population dont le niveau de vie est inférieur au seuil de pauvreté. Ce seuil était de 1063€ par mois en 2018.
- part_pop75 (%): Part des 75 ans ou plus dans la population. Sans faire l'analyse sur les variables, on peut imaginer qu'un département avec une population plus agée pourrait voir son niveau de vie diminuer par exemple. Voilà pourquoi cette indicateur à sa place dans notre étude.
- niveau_vie_median (€): Le niveau de vie est égal aux ressources totales du ménage divisées par le nombre de personnes vivant dans ce ménage. Il détermine l'accès des individus aux biens et services.
- part_20_24_sortis_nondip (%): Rapport de la population des individus âgés de 20 à 24 ans qui ne poursuivent ni études, ni formations.
- part_foy_fisc_impos (%): La part des ménages fiscaux imposés est le pourcentage des ménages fiscaux qui ont un impôt à acquitter au titre de l'impôt sur le revenu. On s'attend à ce que un pourcentage élevé implique un taux de pauvreté plus faible.
- part_enf_famille_ss_empl (%): Une part importante des jeunes de moins de 18 ans vivant dans des ménages sans actif occupé est un signe de précarisation sociale. Nous voulons voir si il y a un lien avec l'indicateur suivant.
- part_jeunes_non_inseres (%): Nous avons voulu mettre en lien cet indicateur avec le précédent et savoir si un pourcentage élevé de famille sans actifs engendre des jeunes ne s'insérant pas professionnellement.
- part_pop_hlm (%): Souvent le lien entre le logement d'une population et son niveau de vie est fait. Nous avons voulu voir si en effet, un grand pourcentage de personne habitant dans un HLM au sein d'un département impliquait un niveau de vie médian plus faible dans celui-ci.
- taux_participation (%): Rapport entre le nombre de votants et le nombre d'inscrits lors du 1er tour des élections législatives de 2017.
- med_dist_min_hp/act (km): Distance entre le domicile et le lieu de travail parcourue pour les 50 % des déplacements domicile/travail les plus courts.
- part_actifs_stables (%): Rapport entre le nombre d'actifs stables et le nombre d'actifs occupés qui résident sur la commune. Un actif stable est un actif qui travaille et réside dans la même commune.

Faisons une rapide étude descriptive de ces 11 variables restantes et regardons la corrélation entre nos variables.

Pour notre étude descriptive, nous allons nous aider des boxplots des variables ainsi que de la commande `summary(res11var[1:96,])`. Nous avons mis ces élèments dans la partie "Annexes" à la fin de notre dossier afin de ne pas polluer visuellement notre rapport. Nous nous aiderons aussi de la matrice de corrélation.

Pour la création de ces boxplots, du sommaire, et de la matrice de corrélation, les DOM ne seront pas pris en compte. Cependant, nous apporterons des informations qui nous ont semblé importantes/pertinantes à leur sujet.

```
corrplot(mat_cor11, type="upper", order="original", tl.col="black", tl.srt=60)
```



Tout d'abord, nous pouvons constater assez surprenamment que la variable 'taux de pauvreté' n'est corrélé significativement qu'avec 2 variables qui sont 'part des 20-24ans non diplômés' et 'part des enfants dans une famille sans actifs occupés', et bien corrélés avec 'niveau_vie_median' et 'part_jeunes_non_inseres'. Nous pouvons voir aussi que la part de la population agée de plus de 75 ans d'un département n'a aucune influence directe (pas de corrélation linéaire) sur le taux de pauvreté. Grâce au `summary`, on notera que le taux de pauvreté médian est de 14,40% en France métropolitaine.

Nous pouvons, avec le boxplot, noter que pour la France métropolitaine, ce taux de pauvreté est assez homogène; il n'y a que les Pyrénées-Orientales et la Seine-saint-Denis qui se démarquent réellement avec un taux de pauvreté supérieure à 20%. Cependant, pour les DOM cela n'a rien à voir. Par exemple, nous pouvons voir que la Réunion a un taux de pauvreté (38,9%) plus de 4 fois supérieurs à la Haute-Savoie(8,9%). De plus, la Réunion n'est certainement pas le département le plus pauvre mais nous n'avons pas souhaité utiliser la valeur de Mayotte étant donné qu'elle a été imputée.

Pour rester sur Mayotte, sa valeur au niveau de la variable 'part_pop75' est assez choquante. En effet, seulement 0,91% de la population a un âge supérieur ou égal à 75 ans. Malheureusement, ceci n'est pas une erreur dans les données puisque l'espérance de vie à la naissance est de seulement 76,3 ans sur cette île contre plus de 80 ans en France métropolitaine.

En valeurs aussi remarquables (d'une moindre mesure), il y a aussi celles de Paris et des Hauts-de-Seine où le niveau de vie médian dépasse les 28000€. Alors que le niveau de vie médian en France est de 21010€. Attention cependant à ne pas tirer de conclusion trop hâtive, ces chiffres ne veulent pas simplement dire que les Parisiens gagnent en moyenne 7000€ de plus par an. En effet, le nombre de personnes au sein des foyers est un facteur important dans le calcul du niveau de vie. Cependant, la variable 'niveau de vie médian' étant fortement corrélés avec la variable 'part de foyer fiscal imposée' ($r > 0.9$), nous pouvons supposer qu'il est plus facile de trouver du travail à Paris et dans les Hauts-de-Seine ce qui peut expliquer en parti ce niveau de vie plus élevé.

Une corrélation assez intéresante est le taux de participation aux élections législatives est qui corrélé avec le taux de personnes ayant plus de 75 ans. Cela veut dire que plus la population d'un département sera agée, plus elle aura tendance à se rendre aux urnes. A l'inverse, on pourrait donc pointer du doigt un désinteressement de la jeunesse pour la politique.

Intéressons nous maintenant aux trois variables touchant les jeunes, la part des 20-24 ans non diplômés, la part des 0-17 ans dans une famille sans actif occupé et la part des 18-25 ans non insérés.

Tout d'abord, nous voyons que ces trois variables sont très fortement corrélées. Il ne semble pas incohérent de dire que des jeunes sans diplômes auront plus de mal à s'insérer professionnellement. Ce qui selon nous est plus intéressant à remarquer, c'est la corrélation entre ces 3 variables avec le taux de pauvreté. Cela veut dire que la pauvreté d'une region a un impact sur la jeunesse. On peut donc se poser la question suivante : La baisse du taux de pauvreté d'un territoire doit-elle passer par un enrichissement éducatif de sa jeunesse ? Malheureusment, nous n'y répondrons pas dans ce rapport.

Pour finir l'étude descriptive, nous allons simplement mentionner le reste des variables corrélées. Nous n'avons pas remarquer de valeurs ou de faits particuliers à nos yeux.

Tout d'abord, on remarque que la variable 'part_pop_hlm' est très corrélée (r < -0,77) avec la part de personnes ayant au moins 75 ans; ce qui semblerait vouloir dire à première vue que les personnes âgées ont tendance à moins vivre dans des HLM. Cependant, nous avons l'impression que cette conclusion est un peu rapide. Ensuite, la distance médiane des navettes domicile-travail pour les actifs est elle aussi fortement corrélées négativement avec la part des personnes ayant plus de 75 ans. Là encore, la liaison ne nous a pas paru très évidente étant donné que les personnes de cet âge ne sont plus comptées dans les actifs (en général). Ce qui paraît plus intéressant est de noter la corrélation entre cette distance et la part de la population vivant dans des HLM.

3. Analyse en composantes principales.

a. Choix du nombre d'axes:

Après cette etude descriptive, nous allons maintenant passer à la partie ACP de notre rapport. Pour cela, on utilise la commande `res.pca <- PCA(res11var[1:96,], graph=FALSE)`. A ce stade, nous sommes obligés de retirer la Guyane, la Réunion et Mayotte car leur contribution aux axes serait beaucoup trop élevée (70% de contribution à eux trois sur la première dimension). Nous aurions aussi pu mettre en individus supplémentaires les DOM mais ils nuieraient un peu à la lisibilité sur les axes plus tard dans notre étude. Nous les rajouterons donc plus tard.

```
res.pca.dt <- PCA(res11var, graph=FALSE)
sort(res.pca.dt$ind$contrib[,1], decreasing = TRUE)[1:10]
```

```
##             Mayotte              Guyane          La Réunion          Guadeloupe
##          41.2135067          23.8651180           7.5849527           4.3413531
##           Martinique      Hauts-de-Seine           Yvelines               Paris
##           3.0413264           2.1485194           1.8340597           1.1954360
## Pyrénées-Orientales        Haute-Savoie
##           0.6626322           0.6505499
```

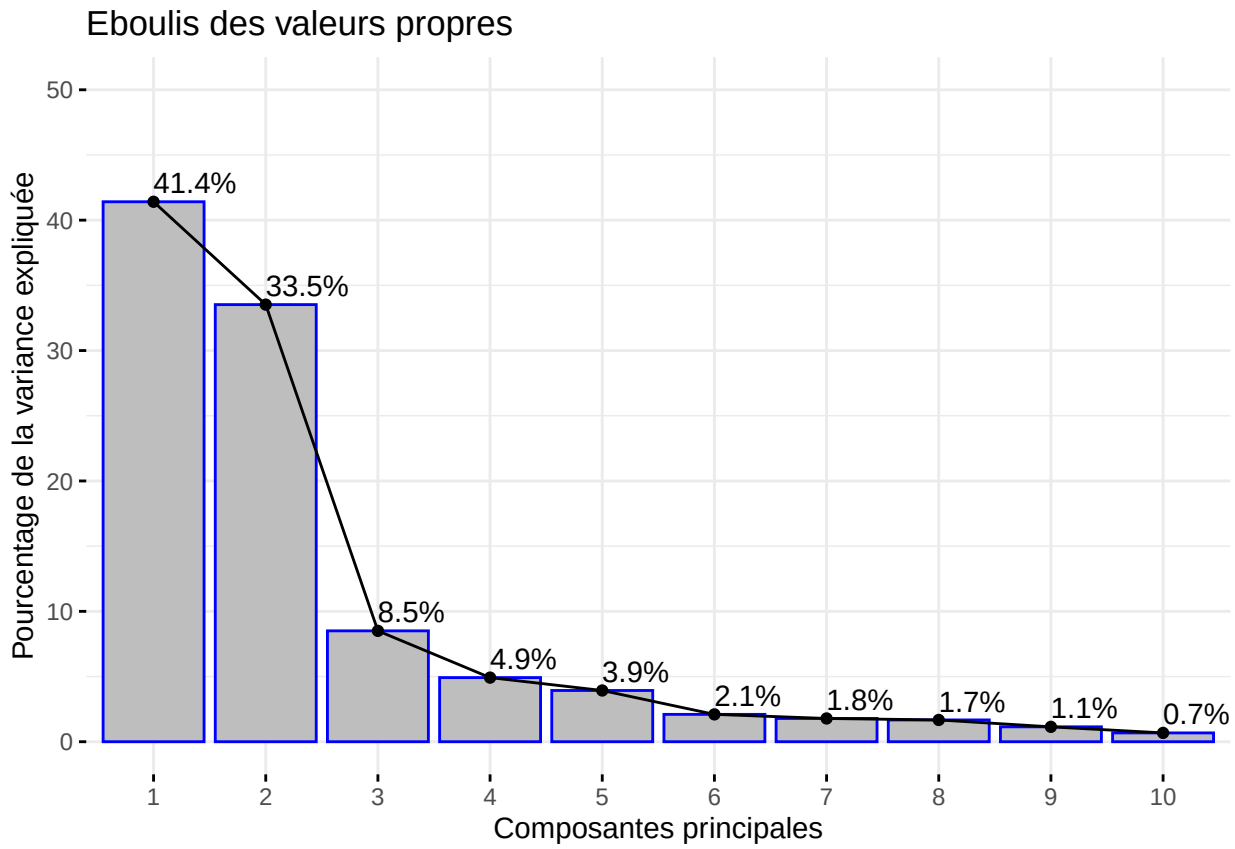
Regardons avant tout combien d'axes principaux nous allons devoir garder.

```
res.pca <- PCA(res11var[1:96,], graph=FALSE)
res.pca$eig
```

```
##         eigenvalue percentage of variance cumulative percentage of variance
## comp 1  4.55514810             41.4104372                           41.41044
## comp 2  3.68761949             33.5238135                           74.93425
## comp 3  0.93540406              8.5036733                           83.43792
## comp 4  0.54113316              4.9193924                           88.35732
```

```
## comp 5  0.43230494              3.9300449                    92.28736
## comp 6  0.23082944              2.0984495                    94.38581
## comp 7  0.19635680              1.7850618                    96.17087
## comp 8  0.18384504              1.6713185                    97.84219
## comp 9  0.12567986              1.1425442                    98.98474
## comp 10 0.07450106              0.6772824                    99.66202
## comp 11 0.03717805              0.3379823                   100.00000
```

```
p = fviz_eig(res.pca, addlabels = TRUE, ylim = c(0,50), barfill = "gray", barcolor = "blue")
p + labs(title = "Eboulis des valeurs propres",
         x = "Composantes principales", y = "Pourcentage de la variance expliquée")
```



En se référant au critère de Kaiser et à la règle du coude, nous devons garder les deux premiers axes principaux. Mais avec seulement deux axes, 75% de l'inertie totale serait expliquée. Nous devons faire un choix. Pour cela, nous allons regarder les variables qui contribuent le plus à la création du troisième axe.

```
sort(res.pca$var$contrib[,3], decreasing = TRUE)
```

```
##      part_actifs_stables           taux_pvt/total      part_foy_fisc_impos
##              67.47855621              10.78028392               6.15654031
##        niveau_vie_median             part_pop_hlm part_enf_famille_ss_empl
##               5.19738199               3.38604675               3.31197691
##  part_jeunes_non_inseres      med_dist_min_hp/act               part_pop75
##               2.72185551               0.85150256               0.06856793
##       taux_participation part_20_24_sortis_nondip
##               0.02722288               0.02006504
```

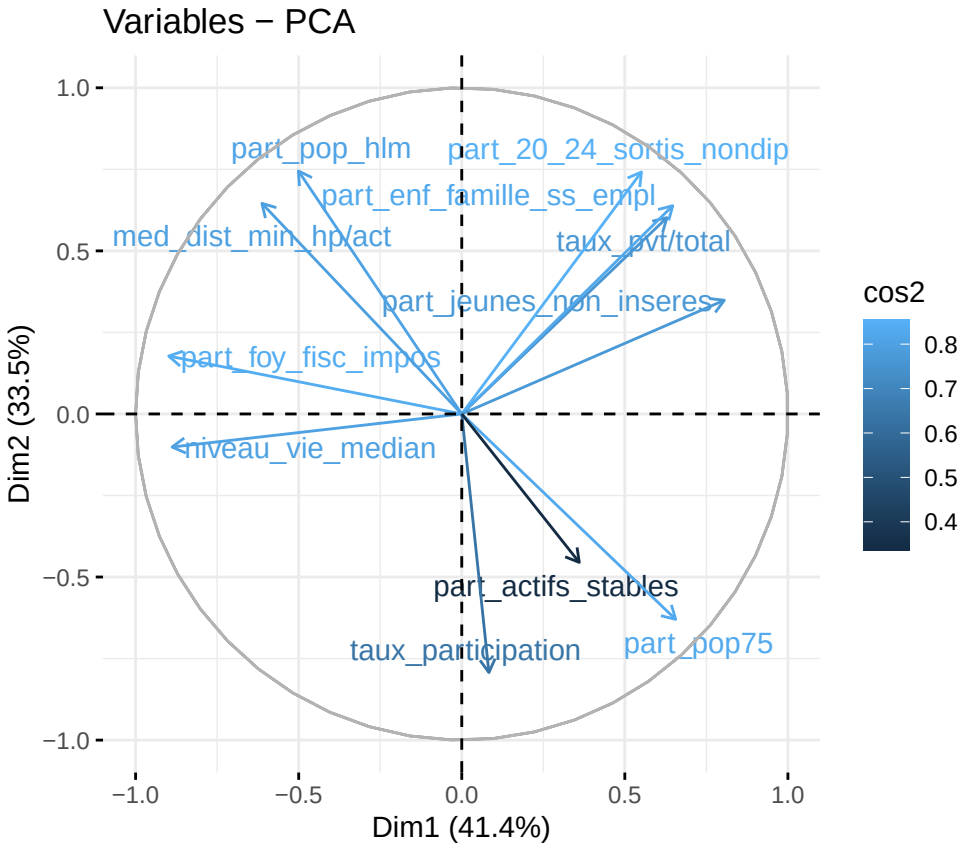
Nous voyons que le troisème axe est créé à 67% par la variable 'part_actifs_stables'. Etant donné qu'elle

ne fait pas partie des variables qui nous intéressent le plus, nous ne garderons que les deux premiers axes principaux.

b. Etude des axes:

Regardons maintenant notre cercle de corrélation :

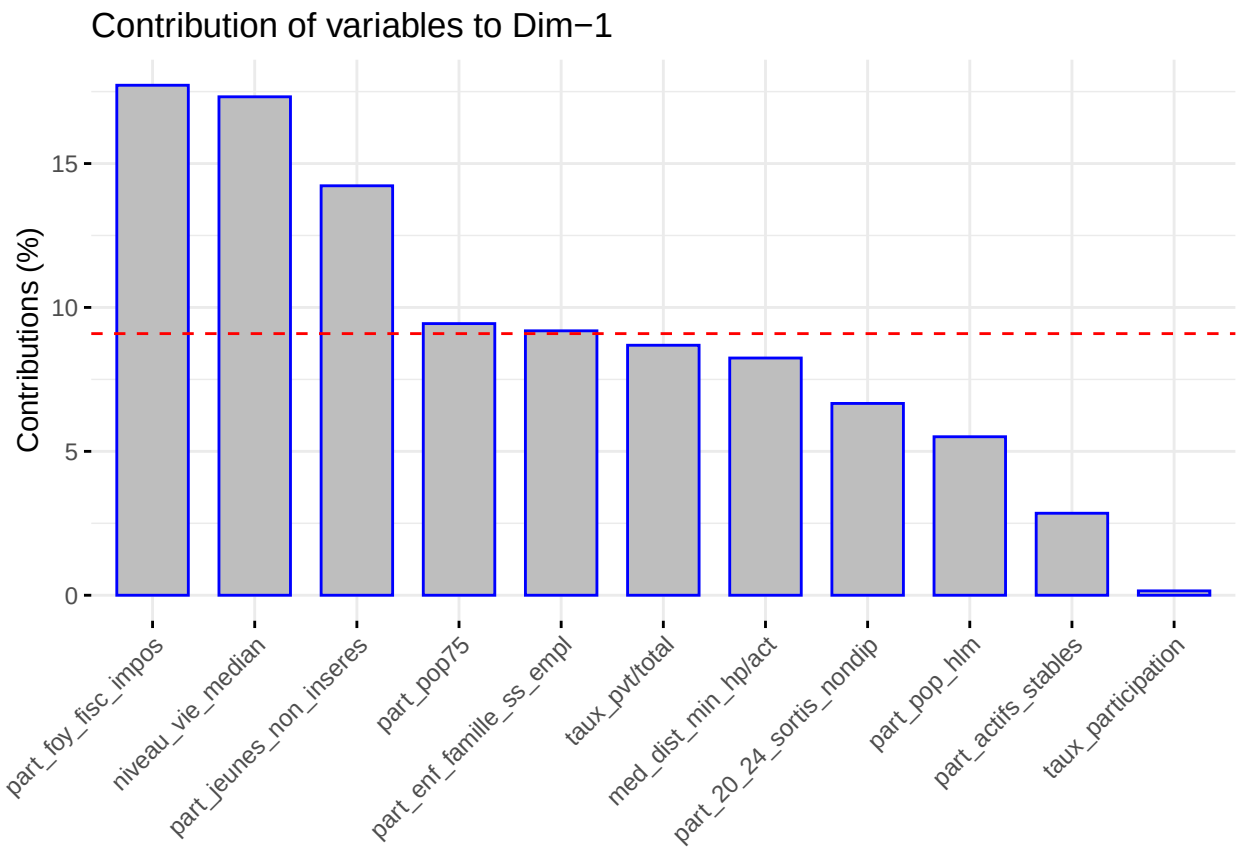
```
fviz_pca_var(res.pca, col.var = "cos2",
             ggtheme = theme_minimal(),
             repel = TRUE
             )
```



Tout d'abord, nous pouvons voir que toutes nos variables sont bien représentées sur le plan principal sauf la part des actifs stables. Ceci est normal, comme nous l'avons dis, la part des actifs stables est beaucoup mieux représentée sur le troisième axe principal.
Nous retrouvons aussi les corrélations énoncées précedemment telles que le taux de pauvreté en corrélation avec les variables touchant la jeunesse ou encore le niveau de vie médian avec la part des foyers fiscaux imposés.

Voyons maintenant queles sont les variables qui contribuent le plus à chacun des axes.

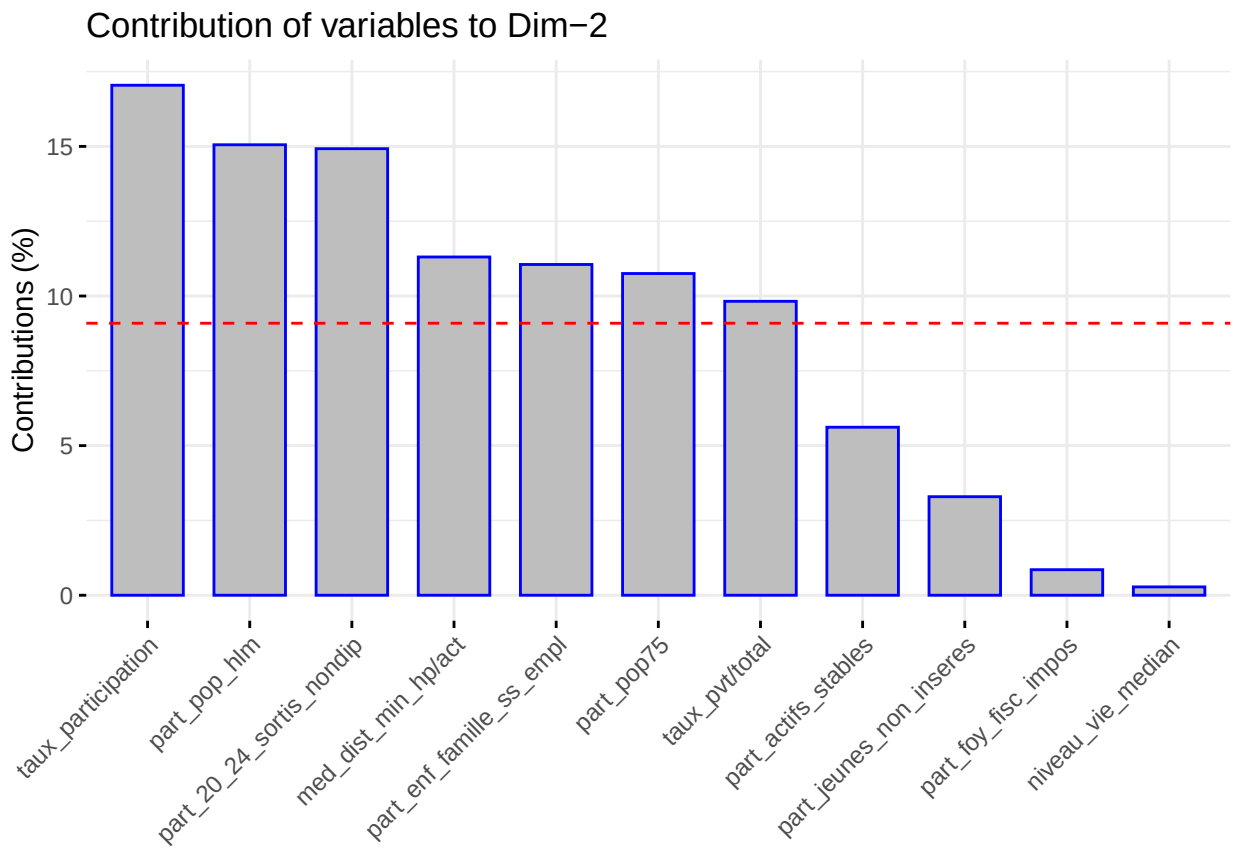
```
fviz_contrib(res.pca, choice = "var", axes = 1, fill = "gray", color = "blue")
```



Avec ce barplot, nous constatons que trois variables se dégagent dans la contribution à la création du premier axe principal. Ces trois variables sont la part des foyers fiscaux imposés, le niveau de vie médian ainsi que la part des jeunes non insérés professionnellement. A première vue, ces variables semblent donner des informations économiques sur les régions.

Intéressons nous maintenant au deuxième axe principal.

```
fviz_contrib(res.pca, choice = "var", axes = 2,, fill = "gray", color = "blue")
```



Passons maintenant au deuxième axe principal.

Grâce à ce barplot, nous voyons que les trois variables contribuant le plus à la création du premier axes sont: le taux de participation, la part de la population habitant dans des HLM et la part des 20-24 ans non diplômés. Trois variables que l'on pourrait qualifier de sociales.

Remarque: La ligne rouge sur les boxplots correspond à la valeur attendue si la contribution des éléments était uniforme.

Avec les deux barplots, nous voyons aussi que le taux de pauvreté contribue assez bien aux deux axes principaux et surtout le deuxième.

On remarque aussi que les trois variables concernant la jeunesse se retrouvent dans le quartan supérieur droit.

Ainsi, un invidu bien représenté sur la droite du premier axe principal devrait faire partie des départements les plus pauvres économiquement de la France métropolitaine. Alors qu'un individu bien représenté sur le haut du deuxième axe principal devrait voir sa jeunesse plutôt défavorisée avec une relative "grande" partie de sa population vivant en HLM et voir sa population moins présente lors des élections.

c. Etudes des individus:

Regardons tout d'abord quels individus contribuent le plus à la création de chacun des deux axes.

```
contr_axe1 <- sort(res.pca$ind$contrib[,1], decreasing = TRUE) ; contr_axe1[1:4]
```

```
##      Hauts-de-Seine           Yvelines              Paris Pyrénées-Orientales
##           10.840080           8.441864           5.826082           5.391258
```

```
contr_axe2 <- sort(res.pca$ind$contrib[,2], decreasing = TRUE) ; contr_axe2[1:4]
```

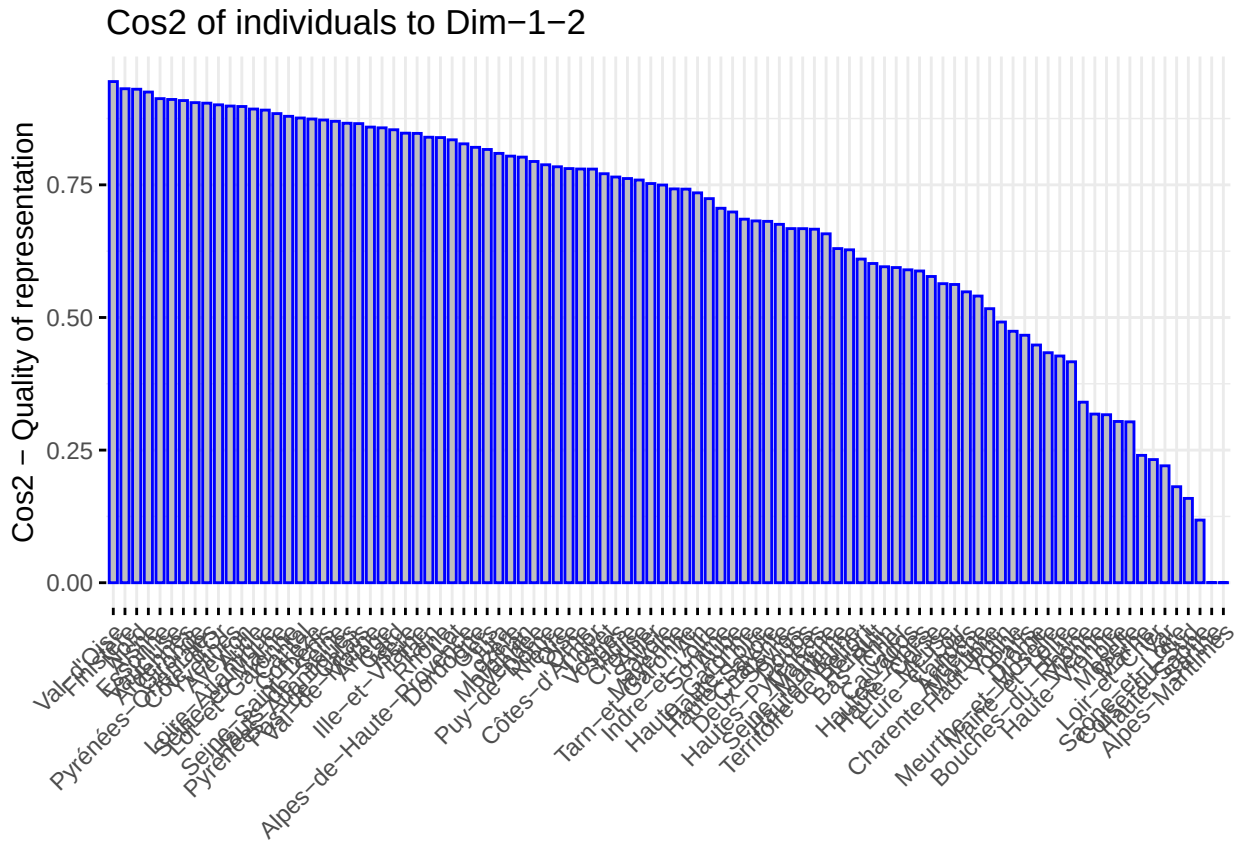
```
## Seine-Saint-Denis        Val-d'Oise             Nord      Val-de-Marne
```

```
##          21.366783          5.411363          4.606843          4.021937
```

Nous remarquons que les individus qui contribuent le plus à la création du premier axe principal sont les Hauts-de-Seine, les Yvelines, Paris et les Pyrénées-Orientales tandis que ceux contribuant le plus à la création du deuxième sont la Seine-Saint-Denis, le Val d'Oise, le Nord et le Val de Marne.

Regardons maintenant si la majorité de nos individus sont bien représentés sur notre plan.

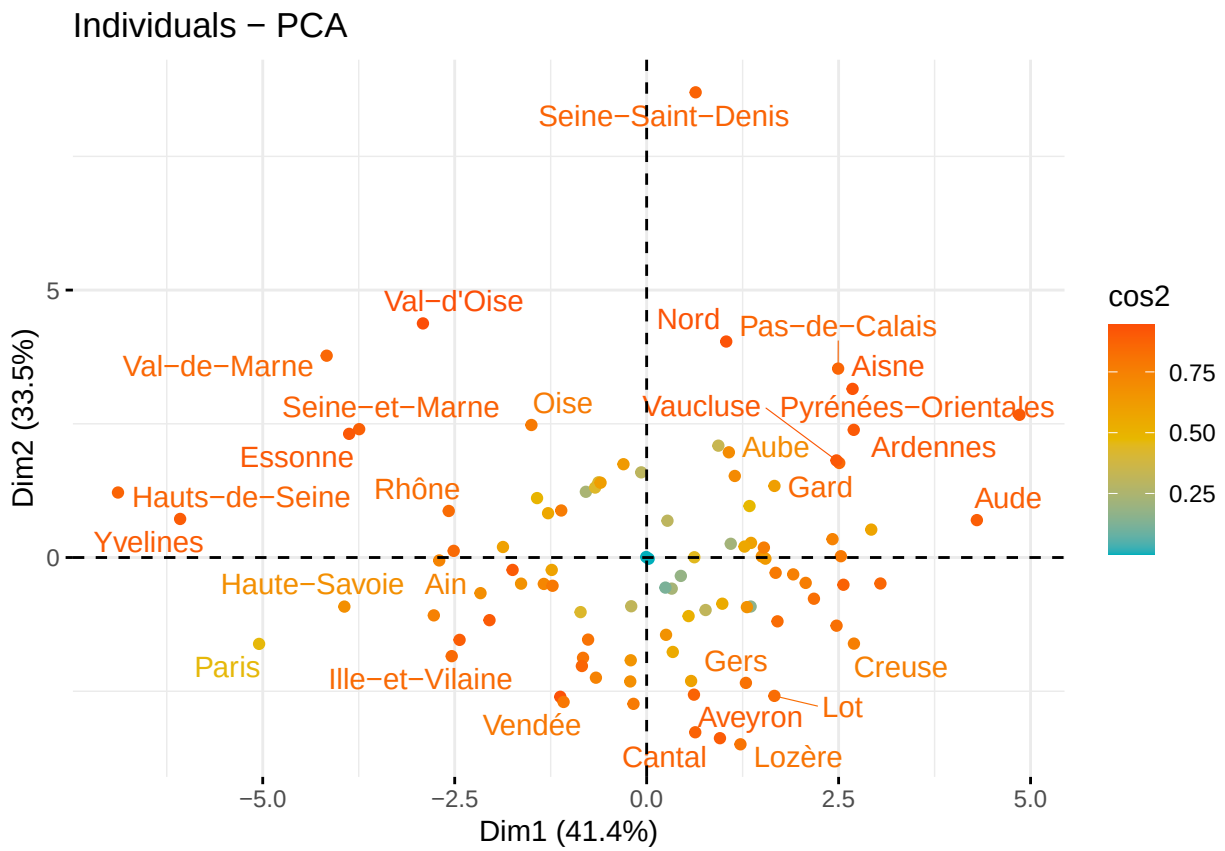
```
fviz_cos2(res.pca, choice = "ind", axes = 1:2,, fill = "gray", color = "blue",tl.cex = 0.5)
```



Bien que ce graphique ne soit que très peu lisible, nous l'admettons, nous pouvons quand même constater que les trois quarts de nos individus ont un cos2 supérieur à 0,5 et que la moitié des individus en ont un supérieur à 0,75, ce qui est plutôt une bonne chose.

```
fviz_pca_ind(res.pca, col.ind = "cos2",
             gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"),
             repel = TRUE
             )
```

```
## Warning: ggrepel: 67 unlabeled data points (too many overlaps). Consider
## increasing max.overlaps
```



A la première vue de ces individus sur le plan principal, la première chose qui nous a sauté au yeux est la position de la Seine-Saint-Denis sur le second axe principal ainsi que celle des Pyrénées-Orientales sur le premier. Notons que ces départements sont bien représentés sur leur axe principal respectif.

```
res.pca$ind$cos2[94,1:2] #qualité de représentation de la Seine-Saint-Denis.
```

```
##       Dim.1       Dim.2
## 0.004669107 0.867584421
```

```
res.pca$ind$cos2[64,1:2] #qualité de représentation des Pyrénées-Orientales.
```

```
##     Dim.1     Dim.2
## 0.6645258 0.1233353
```

La Seine-Saint-Denis est extrêmement haute et très proche du deuxième axe principal. Cette position traduit plusieurs choses et notamment un taux de pauvreté élevé, ainsi qu'une certaine détresse de la jeunesse. En effet, si l'on regarde les valeurs de la Seine-saint-Denis, on remarque que ce département fait partie des cinq départements ayant le niveau de vie médian le plus bas et où la part des jeunes non diplômés est la deuxième plus élevée de la France métropolitaine derrière justement les Pyrénées-Orientales.
Pour les Pyrénées-Orientales, cette position tout à droite du premier axe traduit un taux de pauvreté élevé, un faible taux de foyers imposés ainsi qu'un fort pourcentage de jeunes non insérés professionellement. Comme prévu avec la partie analyse des axes, une position soit 'très haute', soit à 'très à droite' traduit une certaine pauvreté économique du département.

Suite à ça, deux groupes ont attiré notre attention.

Premièrement, celui des départements se trouvant dans le quart inférieur droit du plan, ce sont principalement des régions extrêmement rurales.
Que signifie cette position sur notre plan?

Comme nous l'avons vu avec le biplot, cela veut dire que la part de la population ayant plus de 75 ans est plus élevée dans ces départements. Nous avons donc une population en moyenne plus vieille et forcément vivant moins en HLM, comme nous l'avons vu précedemment avec les corrélations. Aussi, se trouvant sur la droite du plan, ils auront tendance à avoir un taux de pauvreté plus élevé que ceux se situant sur la gauche du plan. A l'opposé, nous retrouvons dans le quart supérieur droit du plan des départements comme le Nord ou le Pas-de-Calais avec une population plus jeune que dans les départements ruraux. Ces départements du Nord et du Pas-de-Calais sont eux aussi parmi les plus pauvres de France avec un taux de pauvreté supérieur à 19%. On rappelle qu'en France métropolitaine, le taux de pauvreté médian est de 14,4%.

```
## [1] "Taux de pauvreté du Nord:"
```

```
## [1] 19.1
```

```
## [1] "Taux de pauvreté du Pas-de-Calais:"
```

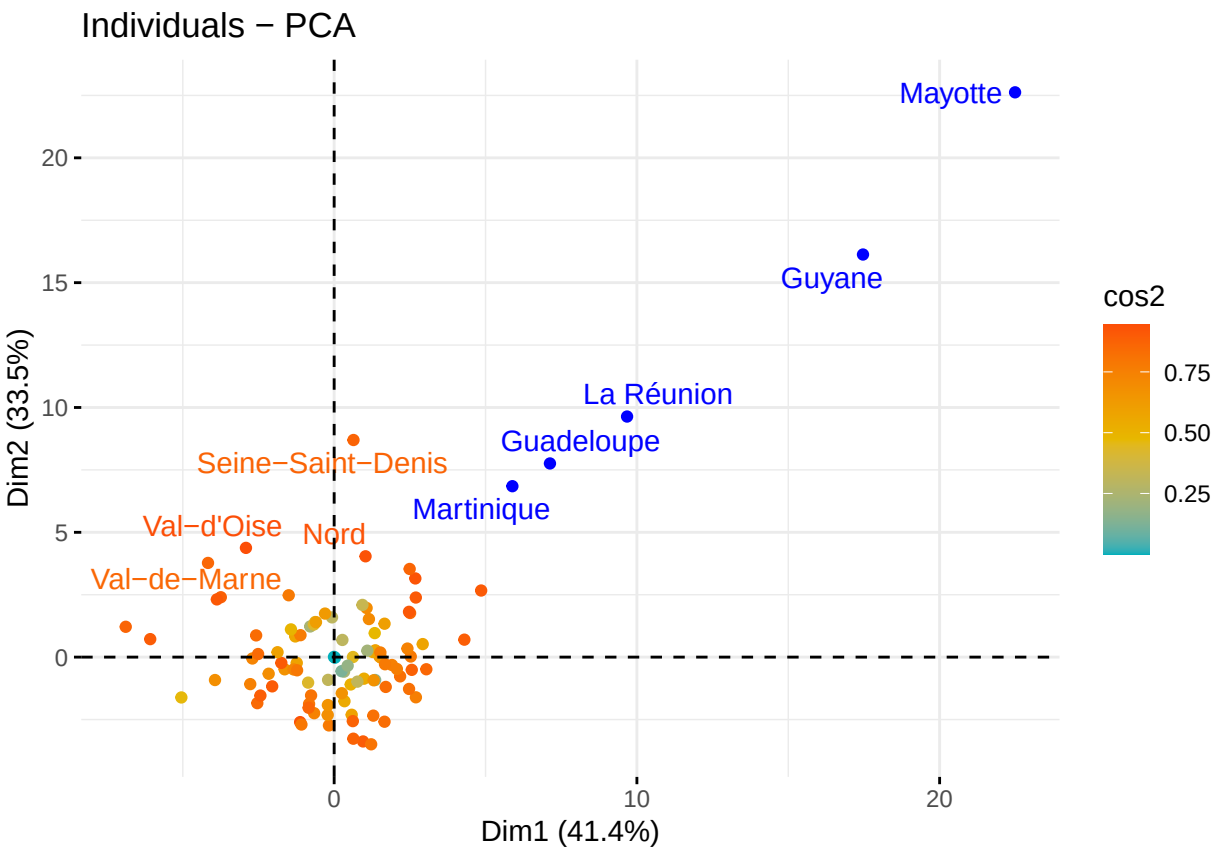
```
## [1] 19.3
```

Deuxièmement, les départements les plus à gauche sont, comme nous l'avons compris, les départements ayant un taux de pauvreté très faible avec un niveau de vie médian élevé ainsi qu'une part des foyers imposés élevés. Ce qui nous a le plus interpellé, c'est le fait que ces départements paraîssent comme détachés des autres. En effet, il semblerait qu'il y ait une différence significative du niveau de vie médian entre les habitants de la Haute-Savoie, de l'Essone, des Yvelines mais surtout des Hauts-de-Seine et Paris avec le reste des départements. Cependant, ce niveau de vie médian élevée ne signifie pas un faible taux de pauvreté. Nous pouvons prendre par exemple Paris et l'Essone qui ne font pas partie des 30 départements ayant les plus faibles taux de pauvreté. Cela veut donc dire qu'il y a de fortes disparités économiques entre certains habitants de ces régions.

Pour finir, nous allons rajouter à notre plan les DOM.

```
fviz_pca_ind(res.pca.dt, col.ind = "cos2",
             gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"),
             repel = TRUE)
```

```
## Warning: ggrepel: 92 unlabeled data points (too many overlaps). Consider
## increasing max.overlaps
```



Voici pourquoi nous n'avons pas fait apparaitre précédemment les DOM.
Leurs valeurs tellement extrêmes empêchent toute lisibilité sur le plan.

Regardons si ils sont bien représentés sur le plan.

```
res.pca.dt$ind.sup$cos2[,1:2]
```

```
##                Dim.1     Dim.2
## Guadeloupe 0.3111648 0.3686019
## Martinique 0.2765777 0.3741766
## Guyane     0.4778570 0.4073395
## La Réunion 0.4061762 0.4030118
## Mayotte    0.4638488 0.4693131
```

Malheureusement, on peut constater qu'ils ne le sont pas vraiment, nous ne pourrons donc pas vraiment continuer l'analyse. Cependant, ces points bien que mal représentés traduisent tout de même une réalité. Se situant au plus profond du quart supérieur droit du plan, les cinq DOM sont les départements ayant le taux de pauvreté le plus élevé, la part des jeunes non diplômés la plus importante ainsi qu'un niveau de vie médian parmis les six plus faibles de France. De plus, avec les taux de participations le plus bas aux legislatives, on pourrait supposer que les habitants ne croient plus aux politiciens pour améliorer leur situation.

4. Régression linéaire.

a. Choix des variables et des individus.

Afin de réaliser notre régression linéaire, nous allons devoir sélectionner les variables qui nous intéressent le plus. Tout d'abord, comme nous n'avons pas parlé de son cas dans la partie ACP, nous ne garderons pas la

variable 'part_actifs_stables'. Aussi, bien qu'arbitraire, il ne nous paraît pas intéressant de garder la variable 'med_dist_min_hp/act'. Nous pouvons donc d'ores et déjà les supprimer de notre tableau 'res11var'.

```
tab_rlm <- res11var[1:96,-c(10,11)]
```

Aussi, comme nous l'avons vu dans la partie ACP, le taux de personne habitant dans les HLM est très corrélé avec le taux de personnes de plus de 75 ans. La part des foyers fiscaux imposés est elle aussi très corrélée avec le niveau de vie médian. On va donc devoir faire attention à ne pas avoir de multicolinéarité dans notre régression. Réalisons notre régression linéaire avec toutes les variables de 'tab_rlm'.

Remarque: Dans cette partie, nous ne prendrons pas en compte les départements d'outre-mer.

```
tab_rlm <- as.matrix(tab_rlm)
```

```
tab_rlm <- as.data.frame(tab_rlm) #lm ne prend ne prend que des dataframes
reg_lin <- lm(`taux_pvt/total` ~ part_enf_famille_ss_empl + part_20_24_sortis_nondip +
                part_jeunes_non_inseres + niveau_vie_median + part_foy_fisc_impos+taux_participation+pa
```

Regardons dès maintenant les facteurs d'influence de la variance (FIV).

```
vif(reg_lin)
```

```
## part_enf_famille_ss_empl part_20_24_sortis_nondip  part_jeunes_non_inseres
##                 3.421892                 5.255804                 4.342180
##        niveau_vie_median      part_foy_fisc_impos       taux_participation
##                 8.874843                10.506927                 2.210832
##             part_pop_hlm               part_pop75
##                 3.842412                 3.588441
```

Comme nous le craignons, la part des foyers fiscaux imposables a un FIV bien trop important. Nous devons donc retirer cette variable de notre régression linéaire.

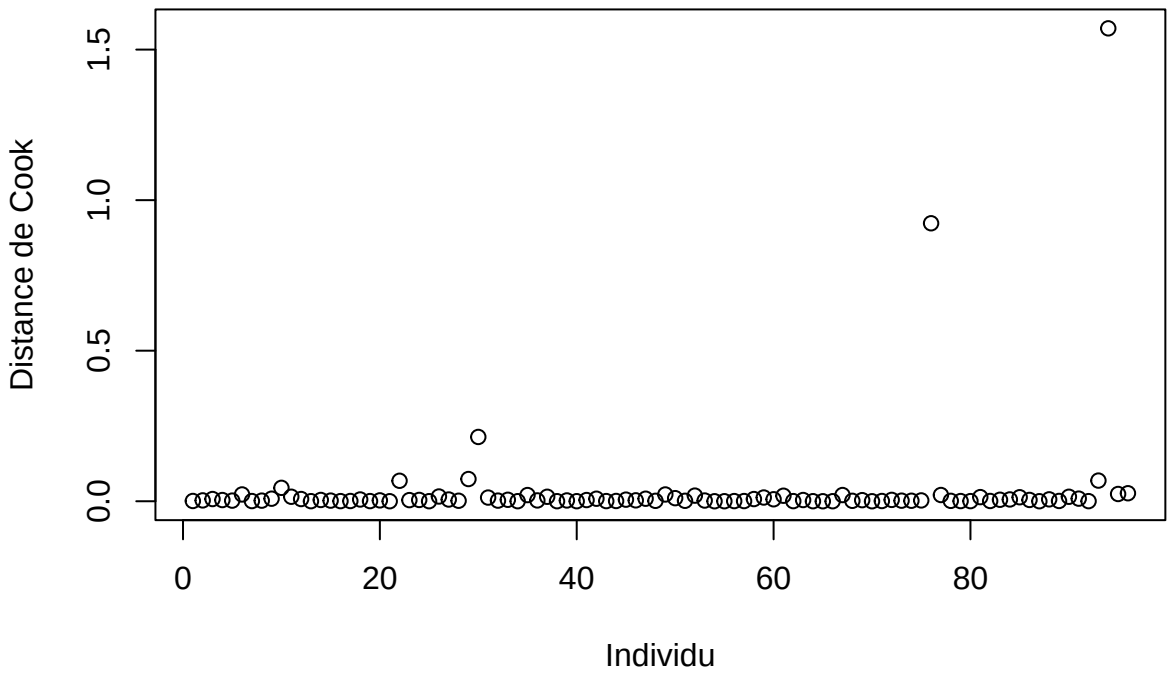
```
reg_lin <- lm(`taux_pvt/total` ~ part_enf_famille_ss_empl
              + part_20_24_sortis_nondip +part_jeunes_non_inseres
              + niveau_vie_median+taux_participation
              +part_pop_hlm+part_pop75 , tab_rlm)
```

Regardons maintenant si un département ne contribue pas de manière abusive à notre modèle. Pour cela, nous utiliserons la distance de Cook.

```
D = cooks.distance(reg_lin)
```

Faisons un plot de ce résultat.

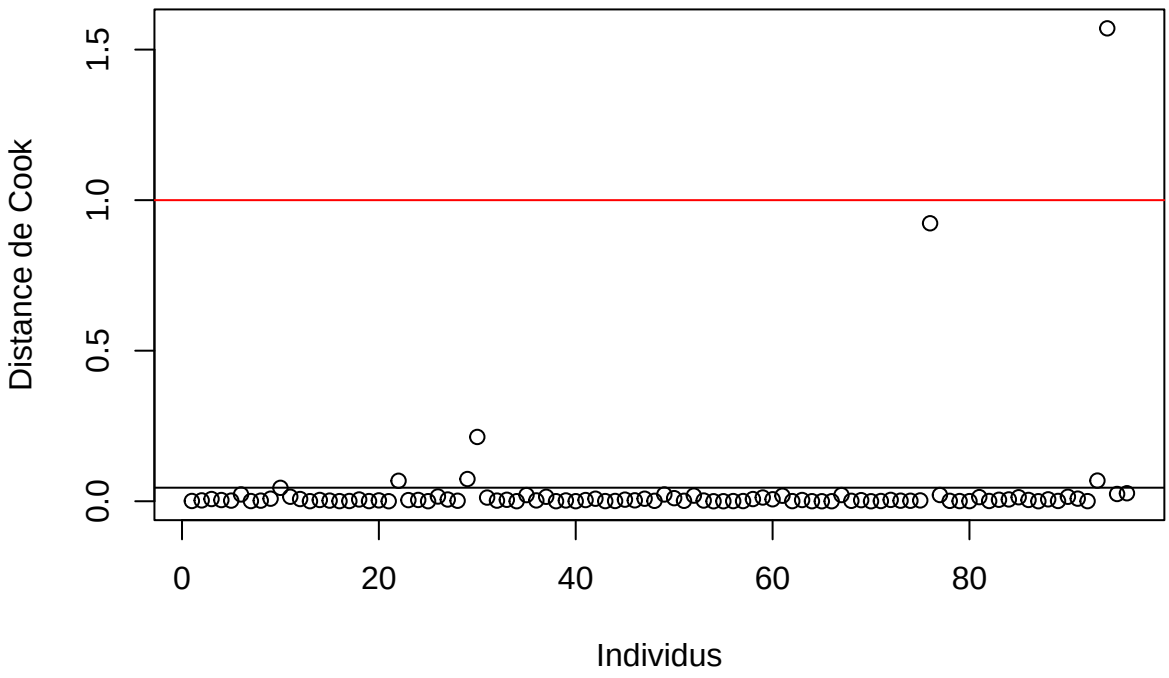
```
plot(D, xlab='Individu',ylab='Distance de Cook')
```



Afin de mieux visualiser, nous allons tracer deux "seuils". Le premier montrera les distances de Cook supérieures à 1 afin de voir si un département a vraiment une influence extrême sur notre modélisation. Toutefois, ce premier seuil est vraiment très permissif. C'est pourquoi le deuxième permettra de détecter si d'autres départements ont des comportements plus influents que les autres mais de façon plus modéré. Ce seuil est 4/(n-p-1) . AVec n le nombre de départements et p le nombre de variables du modèle.

```
{plot(D, xlab='Individus',ylab='Distance de Cook')
abline(1,0,col = 'red')

abline(4/(96-6-1),0)}
```



On voit clairement qu'un département contribue de manière abusive, nous allons donc devoir l'enlever. On constate aussi qu'un autre département a une contribution bien au-dessus du seuil moins permissif. Nous décidons de l'exclure. Pour les quatre autres départements légèrement au-dessus du deuxième seuil, nous faisons le choix de les garder mais il faudra être vigilant.

```
sort(D, decreasing = TRUE)[1:2]
```

```
## Seine-Saint-Denis             Paris
##         1.5705856         0.9231534
```

Nous enlevons donc Paris et la Seine-Saint-Denis.

```
tab_rlm <- as.data.frame(res11var[-c(76,94), ])
```

Nous refaisons notre regression linéaire.

```
reg_lin <- lm(`taux_pvt/total` ~ part_enf_famille_ss_empl
              +part_20_24_sortis_nondip +part_jeunes_non_inseres
              + niveau_vie_median+taux_participation
              +part_pop_hlm+part_pop75 , tab_rlm)
summary(reg_lin)
```

```
##
## Call:
## lm(formula = `taux_pvt/total` ~ part_enf_famille_ss_empl + part_20_24_sortis_nondip +
##     part_jeunes_non_inseres + niveau_vie_median + taux_participation +
##     part_pop_hlm + part_pop75, data = tab_rlm)
##
## Residuals:
```

```
##     Min      1Q  Median      3Q     Max
## -2.5968 -1.0660 -0.2032  0.6627  3.9284
##
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)              17.8868897  5.7000065   3.138 0.002292 **
## part_enf_famille_ss_empl  0.6216094  0.0706705   8.796 8.44e-14 ***
## part_20_24_sortis_nondip  0.2417550  0.0645037   3.748 0.000313 ***
## part_jeunes_non_inseres  -0.0593039  0.0624105  -0.950 0.344516
## niveau_vie_median        -0.0004927  0.0001704  -2.891 0.004801 **
## taux_participation       -0.0387454  0.0495595  -0.782 0.436364
## part_pop_hlm             -0.0241742  0.0468949  -0.515 0.607456
## part_pop75               -0.0007510  0.1163969  -0.006 0.994866
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.476 on 91 degrees of freedom
## Multiple R-squared:  0.9612, Adjusted R-squared:  0.9582
## F-statistic:   322 on 7 and 91 DF,  p-value: < 2.2e-16
```

On voit que les variables "part_pop_75","part_pop_hlm","taux de participation" et "part_jeunes_non_inseres" ont des coefficients très peu significatifs (respectivement 1%, 40%, 57% et 66%). En général, on garde les variables si elles ont un coefficient supérieur à 95%. On va donc encore une fois s'en séparer et nous réalisons donc une nouvelle regression linéaire avec seulement trois variables explicatives.

```
reg_lin <- lm(`taux_pvt/total` ~ part_enf_famille_ss_empl
              +part_20_24_sortis_nondip + niveau_vie_median , tab_rlm)
```

b. Test de significativité et qualité du modèle

Avant d'interpréter les résultats, il nous faut évaluer la significativité du modèle.

Soit l'hypothèse suivante: H0 : absence de significativité globale des variables, i.e au moins une variable n'est pas significativement différente de zéro.

```
summary(reg_lin)
```

```
##
## Call:
## lm(formula = `taux_pvt/total` ~ part_enf_famille_ss_empl + part_20_24_sortis_nondip +
##     niveau_vie_median, data = tab_rlm)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.8053 -1.0864 -0.1073  0.6670  3.9340
##
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)              13.6089925  3.0861754   4.410 2.73e-05 ***
## part_enf_famille_ss_empl  0.6277479  0.0445962  14.076  < 2e-16 ***
## part_20_24_sortis_nondip  0.2176587  0.0503997   4.319 3.86e-05 ***
## niveau_vie_median        -0.0004485  0.0001204  -3.725 0.000332 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.457 on 95 degrees of freedom
```

```
## Multiple R-squared:  0.9605, Adjusted R-squared:  0.9593
## F-statistic: 770.2 on 3 and 95 DF,  p-value: < 2.2e-16
```

On voit que la p-value associée au test de Fischer est extrêmement faible. Ainsi, on peut rejeter fortement H0 et donc le modèle est bien globalement significatif.

Ensuite, nous pouvons aussi voir que le R2 ajusté est supérieur à 0,95. Cela est très correct et donc l'adéquation entre le modèle et les données observées devrait être très forte.

c. Interprétation des résultats.

```
summary(reg_lin)
```

```
##
## Call:
## lm(formula = `taux_pvt/total` ~ part_enf_famille_ss_empl + part_20_24_sortis_nondip +
##     niveau_vie_median, data = tab_rlm)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.8053 -1.0864 -0.1073  0.6670  3.9340
##
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)               13.6089925  3.0861754   4.410 2.73e-05 ***
## part_enf_famille_ss_empl   0.6277479  0.0445962  14.076  < 2e-16 ***
## part_20_24_sortis_nondip   0.2176587  0.0503997   4.319 3.86e-05 ***
## niveau_vie_median         -0.0004485  0.0001204  -3.725 0.000332 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.457 on 95 degrees of freedom
## Multiple R-squared:  0.9605, Adjusted R-squared:  0.9593
## F-statistic: 770.2 on 3 and 95 DF,  p-value: < 2.2e-16
```

En analysant juste les Betas, nous pouvons observer trois choses:
- Baisser de 2% la part de jeunes dans des familles sans actifs baisserait le taux de pauvreté de plus de 1%. (Beta1 vaut 0,63)
- Baisser de 5% la part des 20-24 ans non diplômés baisserait le taux de pauvreté de 1%. (Beta2 vaut 0,22)
- Augmenter significativement le niveau de vie médian (10% par exemple) ne ferait pas reculer le taux de pauvreté de 1%. (Beta3 est quasiment nul, il vaut -0,0004)

Ce dernier point est particulièrement intéressant. Le niveau de vie médian est directement lié au revenu du ménage. Notre modèle nous dis que le niveau de vie médian n'a pas d'impact sur le taux de pauvreté. Cela veut dire qu'augmenter le revenu de chaque ménage ne fera pas baisser le taux de pauvreté ce qui nous a paru à première vue très contre-intuitif. Cela est dû a une erreur de notre part. Après quelques recherches, nous avons vu que le taux de pauvreté était calculé directement à partir du niveau de vie médian. En effet, le taux de pauvreté est le pourcentage de personnes se trouvant sous le seuil de pauvreté. Et ce seuil de pauvreté correspond à 50% du niveau de vie médian.
Pour résumer, si l'on augmente tous les salaires d'un département de 2% par exemple, certes le seuil de pauvreté augmentera, mais le pourcentage de personnes en dessous de ce seuil restera sensiblement le même. Aussi on pourrait se dire que si tous les salaires augmentaient de 2%, cela pourrait être lié à l'inflation. Ainsi, le pouvoir d'achat n'augmenterait pas forcément ce qui impliquerait que les gens ne serait pas forcément "plus riche".
On voit donc qu'augmenter les salaires n'est pas une solution pour faire baisser le taux de pauvreté car les plus pauvres seraient toujours aussi pauvre vis à vis des autres personnes du départements.

Pour faire baisser le taux de pauvreté, il faut donc se concentrer sur les personnes les plus défavorisées et notamment les personnes ayant des enfants de moins de 18 ans et n'ayant pas de travail. Le niveau de vie médian est sensible au nombre de personnes au sein du foyer. Toutes les personnes au sein de celui-ci ont le même niveau de vie et donc les familles avec enfant ont plus d'impact sur le calcul du niveau de vie médian. De ce fait si les parents/tuteurs n'ont pas d'emploi, le foyer fera "plus fortement chuter" le niveau de vie median du département. Ainsi, si l'on augmente uniquement le revenu de ces ménages, en leur trouvant un travail, alors ils devraient passer au-dessus des 50% du niveau de vie médian et donc le taux de pauvreté devrait baisser.

Selon notre analyse, le meilleur moyen de faire baisser le taux de pauvreté est donc de diminuer le taux de chomâge (pour les personnes ayant déjà travaillées ce qui semble plus correspondre à des adultes ayant des enfants) et de faciliter l'insertion des jeunes dans la vie active comme nous l'avons vu dans la partie ACP. En effet, nous avons vu que le taux de pauvreté était aussi corrélé avec le taux de jeunes non insérés professionellement.

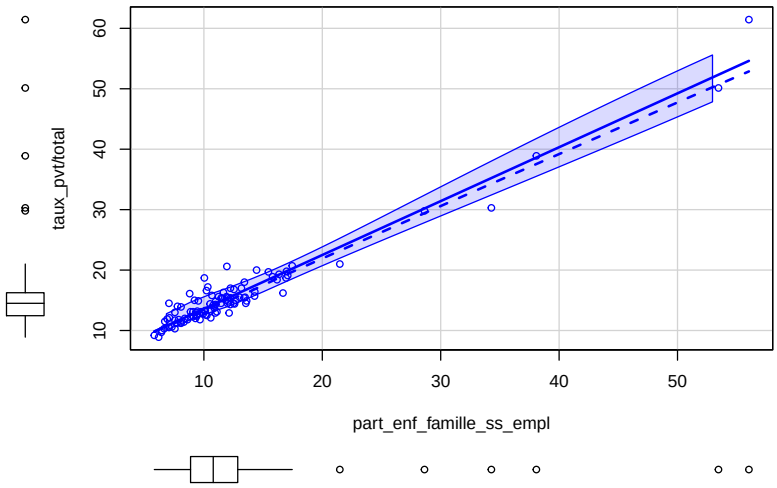
Afin d'illustrer ceci, réalisons une simulation en baissant le taux de jeunes dans des familles sans emploi pour voir l'impact sur le taux de pauvreté.
Prenons pour notre illustration les Pyrénées-Orientales qui comme nous l'avons vu précédemment est un département très pauvre avec un taux de jeunes dans des familles sans emploi parmi le plus élevé de la France métropolitaine.

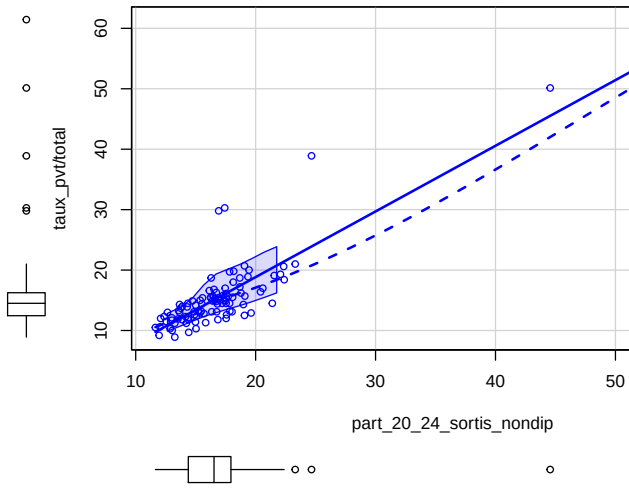
```
predict(reg_lin, newdata=data.frame(part_enf_famille_ss_empl=19,48,
                                    part_20_24_sortis_nondip=23.3,
                                    niveau_vie_median=19350.00),
        se.fit=TRUE, interval = "prediction", level = 0.99)
```

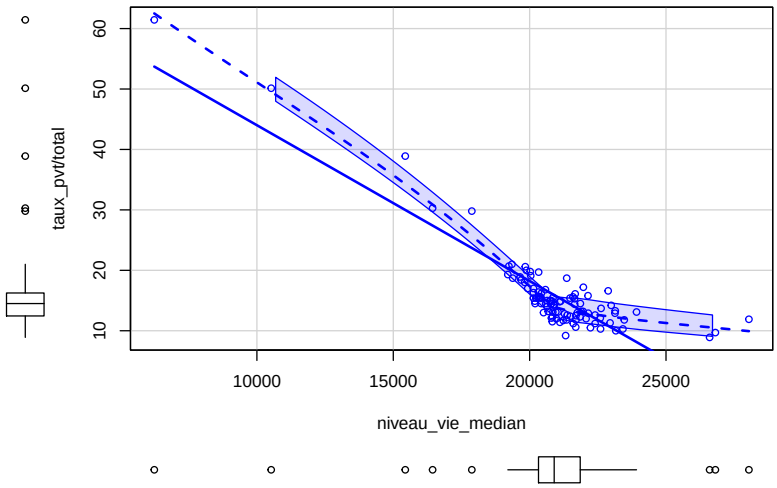
```
## $fit
##        fit      lwr      upr
## 1 21.92987 18.05736 25.80238
##
## $se.fit
## [1] 0.2162315
##
## $df
## [1] 95
##
## $residual.scale
## [1] 1.45728
```

Nous n'avons modifié que la valeur de "part_enf_famille_ss_empl". Nous l'avons baissé de 2%. Initialement, le taux de pauvreté était de 21%.
Selon notre modèle, si l'on baisse la part des enfants dans une famille sans emploi, alors le taux de pauvreté des Pyrénées-Orientales devrait être compris entre 18% et 25%. L'intervalle est assez large mais nous pouvons évidemment enlever les valeurs predictives qui prévoient une augmentation du taux de pauvreté.
Ainsi, en baissant la part des jeunes avec des parents sans emploi de 2%, cela pourrait produire une baisse jusqu'à 3% du taux de pauvreté.
Biensûr, nous savons que cette prédiction ne prend pas en compte tous les autres indicateurs et c'est pourquoi elle ne nous permet que de donner une idée de ce que serait le taux de pauvreté sous ces conditions.

```
scatterplot(`taux_pvt/total`~part_enf_famille_ss_empl, tab_rlm)
scatterplot(`taux_pvt/total`~part_20_24_sortis_nondip, tab_rlm)
scatterplot(`taux_pvt/total`~niveau_vie_median, tab_rlm)
```



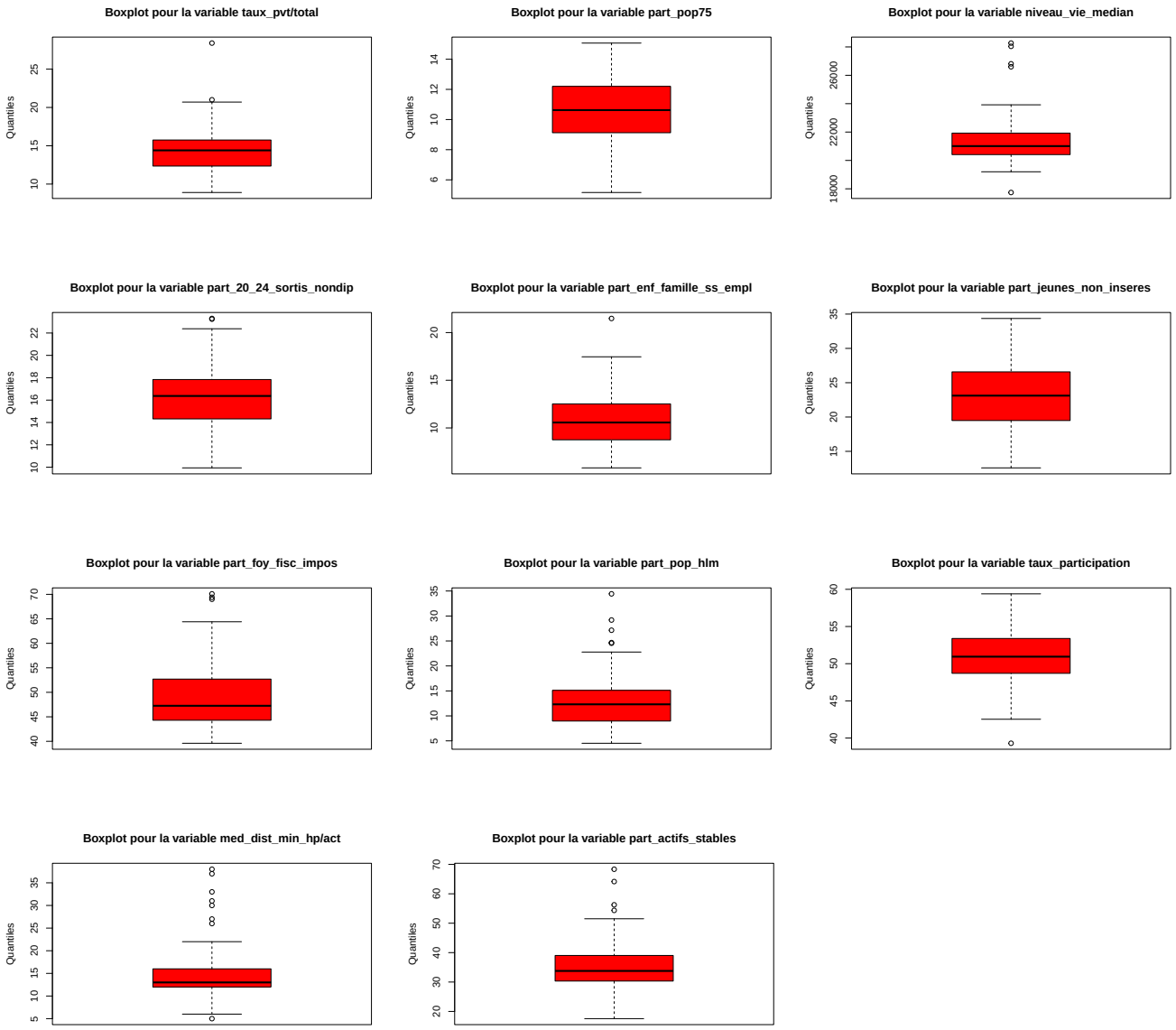




Voici les courbes de régression de nos trois variables explicatives du taux de pauvreté.
On se rend compte que l'hypothèse de linéarité est notamment satisfaite pour la part des enfants dans des familles sans emplois. Notre analyse précédente a donc de grande chance d'être correcte.

5. Annexes.

```
for (i in 1:ncol(res11var)){
  boxplot(res11var[1:96,i],
        col = c("red"),
        main = paste("Boxplot pour la variable",colnames(res11var)[i]),
        ylab = "Quantiles")
  }
```




```
summary(res11var[1:96,])
```

```
##  taux_pvt/total    part_pop75     niveau_vie_median part_20_24_sortis_nondip
##  Min.   : 8.90   Min.   : 5.160   Min.   :17740     Min.   : 9.94
##  1st Qu.:12.38   1st Qu.: 9.143   1st Qu.:20415     1st Qu.:14.34
##  Median :14.40   Median :10.625   Median :21010     Median :16.37
##  Mean   :14.53   Mean   :10.568   Mean   :21395     Mean   :16.35
##  3rd Qu.:15.72   3rd Qu.:12.195   3rd Qu.:21908     3rd Qu.:17.84
##  Max.   :28.40   Max.   :15.070   Max.   :28270     Max.   :23.30
##  part_enf_famille_ss_empl part_jeunes_non_inseres part_foy_fisc_impos
##  Min.   : 5.810           Min.   :12.59           Min.   :39.60
##  1st Qu.: 8.777           1st Qu.:19.50           1st Qu.:44.35
##  Median :10.570           Median :23.11           Median :47.25
##  Mean   :10.926           Mean   :23.05           Mean   :48.91
##  3rd Qu.:12.510           3rd Qu.:26.55           3rd Qu.:52.65
```

```
##  Max.   :21.480           Max.   :34.36           Max.   :70.10
##   part_pop_hlm    taux_participation med_dist_min_hp/act part_actifs_stables
##  Min.   : 4.520   Min.   :39.30      Min.   : 5.00       Min.   :17.55
##  1st Qu.: 8.975   1st Qu.:48.81      1st Qu.:12.00       1st Qu.:30.34
##  Median :12.315   Median :50.95      Median :13.00       Median :33.77
##  Mean   :12.901   Mean   :51.11      Mean   :14.76       Mean   :34.46
##  3rd Qu.:15.102   3rd Qu.:53.38      3rd Qu.:16.00       3rd Qu.:38.97
##  Max.   :34.430   Max.   :59.39      Max.   :38.00       Max.   :68.35
```

```
summary(res11var[97:101,])
```

```
##  taux_pvt/total    part_pop75    niveau_vie_median part_20_24_sortis_nondip
##  Min.   :29.80   Min.   :0.910   Min.   : 6237     Min.   :16.92
##  1st Qu.:30.29   1st Qu.:1.880   1st Qu.:10524     1st Qu.:17.42
##  Median :38.90   Median :4.710   Median :15440     Median :24.66
##  Mean   :42.11   Mean   :5.074   Mean   :13305     Mean   :32.96
##  3rd Qu.:50.13   3rd Qu.:8.260   3rd Qu.:16443     3rd Qu.:44.56
##  Max.   :61.44   Max.   :9.610   Max.   :17880     Max.   :61.22
##  part_enf_famille_ss_empl part_jeunes_non_inseres part_foy_fisc_impos
##  Min.   :28.63            Min.   :34.92           Min.   : 0.4353
##  1st Qu.:34.28            1st Qu.:37.38           1st Qu.:14.7601
##  Median :38.08            Median :42.57           Median :28.3000
##  Mean   :42.10            Mean   :46.05           Mean   :22.1029
##  3rd Qu.:53.46            3rd Qu.:52.57           3rd Qu.:33.2193
##  Max.   :56.03            Max.   :62.80           Max.   :33.8000
##   part_pop_hlm   taux_participation med_dist_min_hp/act part_actifs_stables
##  Min.   :15.49   Min.   :11.97      Min.   : 0.0        Min.   :41.96
##  1st Qu.:15.87   1st Qu.:24.80      1st Qu.: 0.0        1st Qu.:42.79
##  Median :16.88   Median :25.61      Median : 9.0        Median :52.38
##  Mean   :17.38   Mean   :24.63      Mean   : 6.8        Mean   :52.50
##  3rd Qu.:17.36   3rd Qu.:26.08      3rd Qu.:12.0        3rd Qu.:59.46
##  Max.   :21.29   Max.   :34.69      Max.   :13.0        Max.   :65.93
```