# Etude des violences sexuellles par département en France

Bill Yehouenou, Jeanne Ropert

2023-01-19

## Contents



## Problématique

Les violences sexuelles se définissent comme « *toute atteinte sexuelle commise sans le consentement d'une personne ou tout agissement discriminatoire fondé sur le sexe* ». L'évolution des mentalités et de la loi a permis ces dernières années, une libération de la parole. En effet, le nombre de violences sexuelles recensées est en hausse de 33% en 2021. De plus, selon une enquête publiée en 2021 par l'INSERM, « **14,5 % des femmes et 6,4 % des hommes en France, soit environ 5,5 millions de personnes, auraient été confrontés avant l'âge de 18 ans à des violences sexuelles** ». Cette dernière nous apprend également que ces violences apparaissent dans la majeure partie des cas dans un cadre familial. Nous nous sommes questionnées sur les facteurs qui pourraient influencer le nombre de violences sexuelles, en analysant les violences sexuelles physiques par département de France métropolitaine en 2018. En effet, **comment**

Table 1: Dictionnaire des variables

| Code | Définition | Nature | Signe attendu |
|---|---|---|---|
| Faits | Mediane du revenu disponible | Quantitative | + |
| Mediane_revenu_dispo | Population totale du département | Quantitative | + |
| Homme_sans_diplome | Effectif des hommes sans diplome | Quantitative | + |
| Femme_sans_diplome | Effectif des femmes sans diplome | Quantitative | + |
| Taux_chomage | Taux de chômage en % | Quantitative | + |
| Taux_pauvrete | Taux de pauvrete en % | Quantitative | + |
| Taux_logements_sociaux | Taux de logements sociaux en % | Quantitative | + |
| Sexe_politique | Sexe personalité politique | Indicatrice | +/- selon le cas |
| Geographie | Situation géographie | Indicatrice | +/- selon le cas |
| Nombre de faits de violences pour 1000 | Quantitative | | Faits |

**expliquer les différences du nombre de faits de violences sexuelles entre les départements ?**. Ainsi, l'équation de notre modèle devrait se présenter comme suit :

$$Faits = b_0 + b_1 * Mediane.revenu + b_2 * Pop + b_3 * Homme.sans.diplome + b_4 * Femme.sans.diplome +$$

$$b_5 * Taux.chomage + b_6 * Taux.pauvrete + b_7 * Taux.logements.sociaux + b_8 * Sexe.politique + b_9 * Geographie + e$$

# Collecte des données

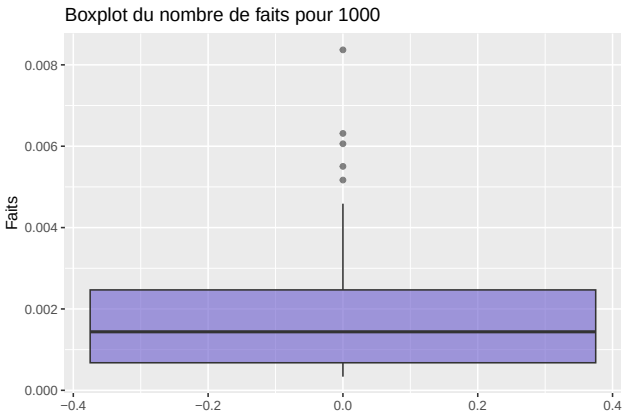
Pour commencer notre étude, nous avons constitué notre base de données, pour cela nous avons sélectionné neuf variables explicatives et une variable à expliquer. Notre variable endogène est donc le *nombre de faits de violences sexuelles physiques par département.*

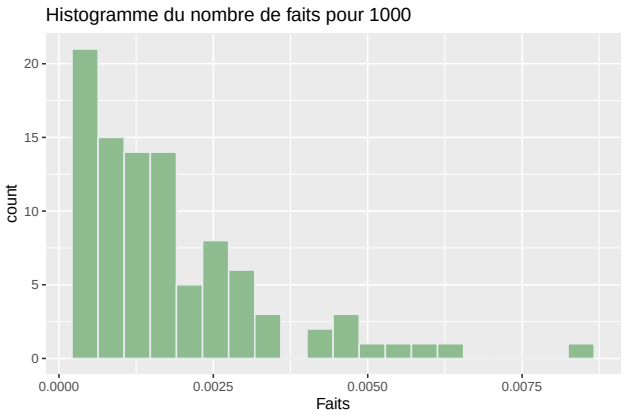
Concernant, nos variables exogènes, nous nous sommes d'abord intéressées au niveau de vie des départements. Pour cela, nous avons réuni des données concernant *le revenu médian disponible en euros*, *le taux de pauvrete*, *le taux de chômage* et *le taux de logements sociaux.* Également, nous avons voulu faire intervenir quelques caractéristiques sur la population, comme son nombre ou encore les effectifs des hommes et femmes de plus de 25 ans sans diplômes. Pour finir, nous nous sommes penchées sur deux caractéristiques des départements, notamment *la situation géographique*, codée 0 pour le nord et 1 pour le sud, et *le sexe du président du conseil départemental*, codé 0 pour les hommes et 1 pour les femmes. Nous avons recueilli ces données concernant 96 départements sur le site du gouvernement, de l'Insee et de l'observatoire des territoires.

# Partie 1 : Etude de l'hétéroscédasticité

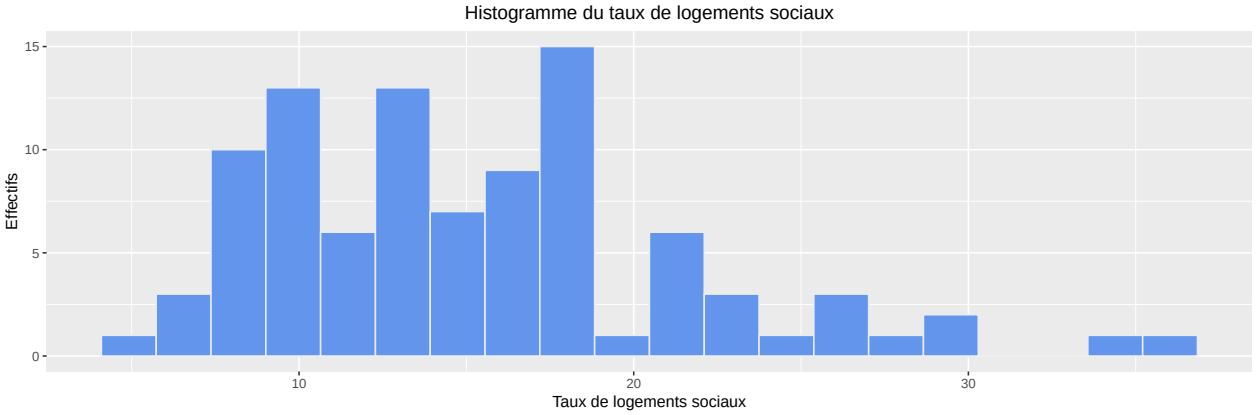
## Statistiques descriptives univariées

Pour en apprendre davantage sur notre variable d'intérêt, nous allons procéder à une analyse statistique desctriptive. La boîte à moustache de notre variable d'intérêt représente une **variabilité assez faible** mais avec des valeurs atypiques notables.



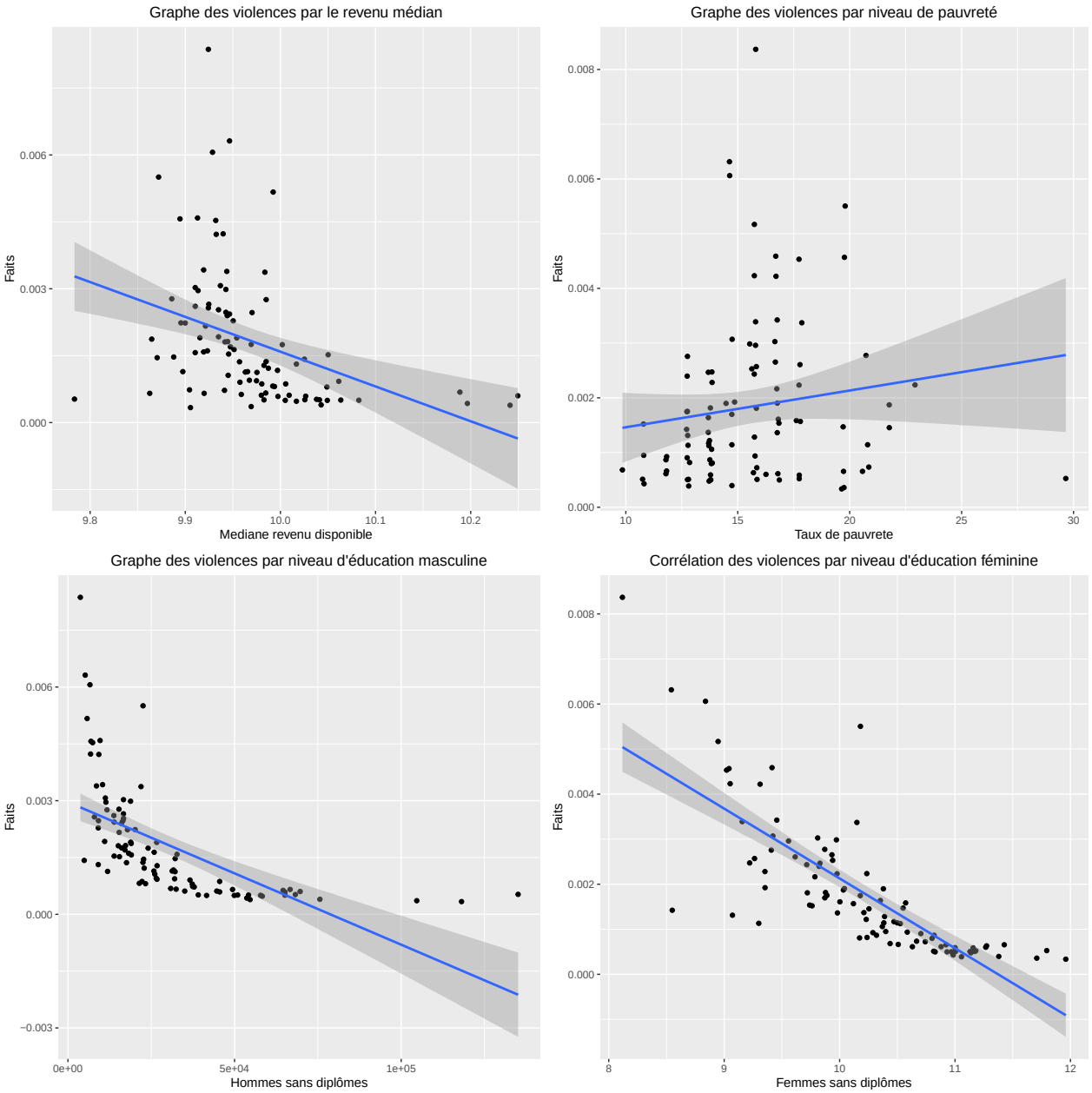

L'analyse de l'histogramme du nombre de faits montre que les observations sont assez concentrées entre 0 et 1 fait pour 1000 habitant. Toutefois, elle possède quelques outliers qu'il faudra surveiller. Nous avons décidé de nous intéresser à la variable *taux de logements sociaux* qui, d'àprès notre revue de littérature, est une variable assez importante.




L'histogramme du *taux de logements sociaux* révèle des valeurs atypiques. De plus, on constate que plusieurs départements sont en dessous du quota de logements sociaux prévus en France, soit 10%. En termes de distribution, la variable du *taux de logement sociaux*, sa distribution semble asymétrique ce qui peut présenter un problème pour notre modèle.

## Statistiques descriptives bivariées

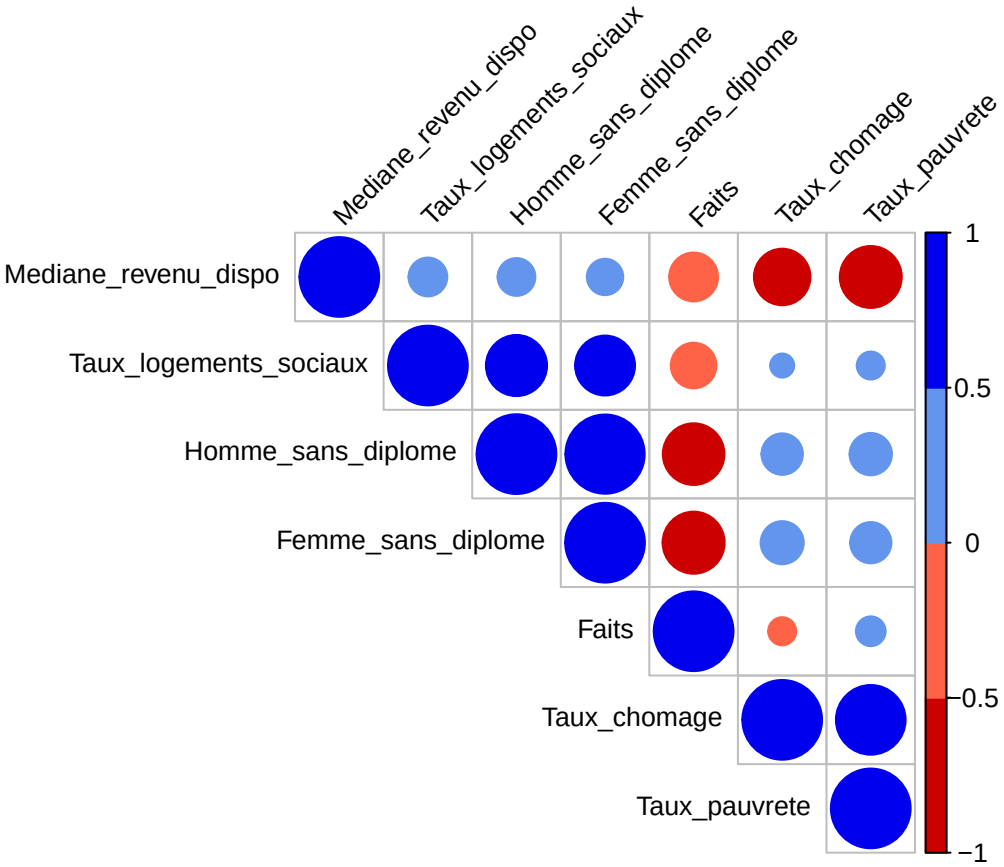
Pour en apprendre davantage sur le type de lien qui existe entre notre variable endogène et nos variables exogènes, nous utilisons la représentation graphique en nuage de points. Tout d'abord, comme nous le montre les graphiques ci-dessous, la médiane du revenu disponible, le taux de pauvrete et le nombre d'hommes et femmes sans diplômes sont négativement et linéairement corrélés à notre variable endogène, c'est-à-dire le nombre de violences sexuelles.




Ainsi, cette première partie nous a permis de visualiser l'ensemble de notre base de données, de constater les éventuels problèmes, les corrélations et le type de lien entre les variables. Cette étape est nécessaire pour poursuivre au mieux la réalisation du modèle économétrique.

# Matrice de corrélation

Désormais, nous allons procéder à une analyse statistique descriptive bivariée afin d'étudier les relations entre le nombre de violences sexuelles recensées par département et nos variables explicatives, mais également les possibles liens entre nos variables explicatives elles-mêmes.




De notre côté, les faits de violences sexuelles sont fortement corrélés au nombre de femmes et d'hommes de plus de 25 ans sans diplômes, et plus légèrement corrélés à la médiane du revenu disponible et au taux de logements sociaux. Par ailleurs, cette matrice met en évidence des corrélations entre des variables explicatives. En effet, les variables Homme et Femme sans diplômes ou encore Revenu disponible et Taux de pauvreté sont corrélées entre elles. Une présence de colinéarité entre des variables explicatives peut altérer notre modèle, un choix devra donc être fait au moment de la régression.

## Modèle

### Spécification du modèle

Afin d'estimer le nombre de faits de violences sexuelles par département, nous devons mettre en place un modèle économétrique constitué des variables qui pourraient expliquer ces violences. Pour choisir le modèle qui estimera le mieux les violences sexuelles, nous commençons par effectuer une régression linéaire du nombre de faits par département, en fonction des neuf autres variables

```
================================================================================
                                        Dependent variable:
                    ----------------------------------------------------------
                                              Faits
                           (1)                (2)                (3)
--------------------------------------------------------------------------------
Mediane_revenu_dispo      -0.00000          -0.00000***
                          (0.00000)          (0.00000)

Homme_sans_diplome        -0.00000*
                          (0.00000)
```

| | | | |
|---|---|---|---|
| Femme_sans_diplome | 0.000 | | |
| | (0.00000) | | |
| Taux_chomage | -0.0004*** | -0.001*** | -0.0005*** |
| | (0.0001) | (0.0001) | (0.0001) |
| Taux_pauvrete | 0.0003*** | 0.0002*** | 0.0003*** |
| | (0.0001) | (0.0001) | (0.0001) |
| Taux_logements_sociaux | 0.00004 | -0.0001** | -0.0001*** |
| | (0.00003) | (0.00002) | (0.00003) |
| Sexe_politique1 | 0.0005 | | |
| | (0.0003) | | |
| Geographie1 | 0.0003 | | -0.0001 |
| | (0.0003) | | (0.0003) |
| Constant | 0.006** | 0.012*** | 0.003*** |
| | (0.003) | (0.003) | (0.001) |
| Observations | 96 | 96 | 96 |
| R2 | 0.576 | 0.349 | 0.271 |
| Adjusted R2 | 0.537 | 0.321 | 0.239 |
| Residual Std. Error | 0.001 (df = 87) | 0.001 (df = 91) | 0.001 (df = 91) |
| F Statistic | 14.766*** (df = 8; 87) | 12.206*** (df = 4; 91) | 8.460*** (df = 4; 91) |
| Note: | | | *p<0.1; **p<0.05; ***p<0.01 |

Après avoir mis en route et comparé différents modèles, nous avons retenu un modèle (1) avec toutes les variables. En effet, il s'agit du modèle avec le meilleur R2 ajusté. Il faudra faire un test de Ramsey pour vérifier la spécification du modèle.

**Test de Ramsey**

```
    RESET test

data:  model1
RESET = 25.57, df1 = 2, df2 = 85, p-value = 2.023e-09
```

La p-value du test de Ramsey est inférieure au seuil de 5% alors, le modèle est bien spécifié.

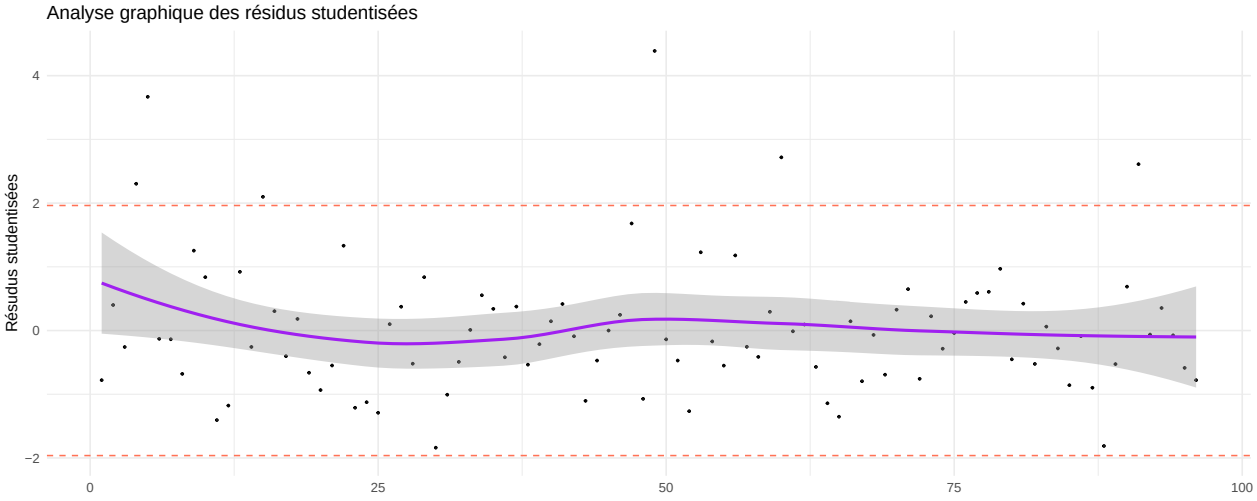
**Analyse des résidus du modèle**

**Résidus studentisés**

Les résidus studentisés obtenus par validation croisée nous montrent que les départements cités ci-dessous sont des départements atypiques dans notre base de données.

```
[1] "Alpes-de-Haute-Provence" "Hautes-Alpes"
[3] "Cantal"                  "Lozère"
[5] "Nord"                    "Territoire de Belfort"
```

On a procédé aussi à un lissage de nos résidus et ce lissage est assez proche de 0. Donc, on soupçonne que nos résidus ont une espérance nulle et de variance constante.

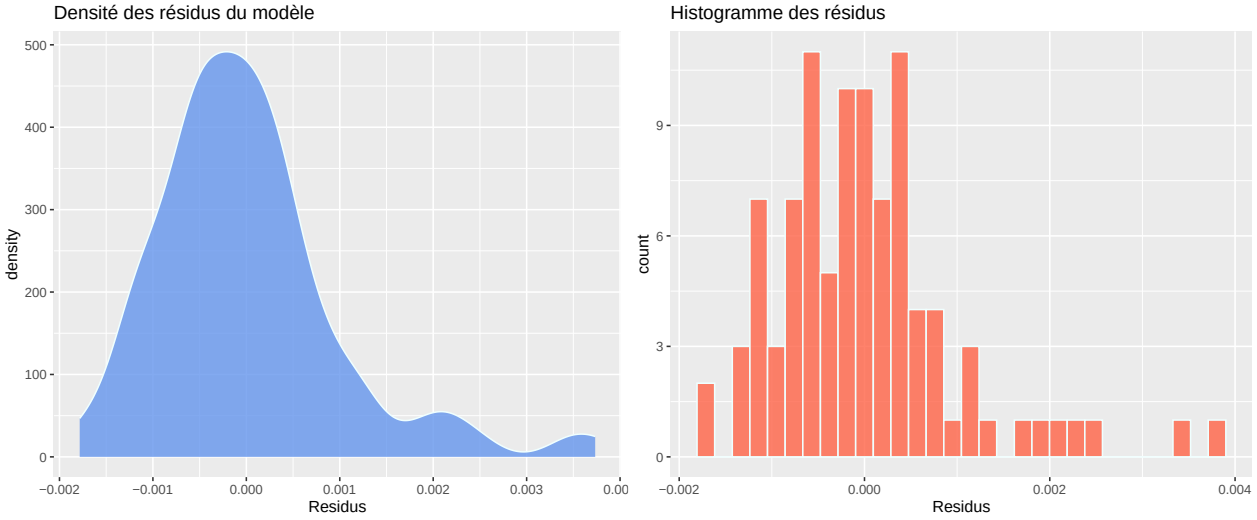



**Test de normalité des résidus de Shapiro-Wilk**

```
    Shapiro-Wilk normality test

data:  model1$residuals
W = 0.91549, p-value = 1.197e-05
```

La p-value du test de normalité des résidus est inférieure au seuil de 5% alors on rejette l'hypothèse de normalité des résidus. Les résidus ne suivent pas une loi normale. Cela est dû à l'existence de valeurs atypiques qui étendent la distribution.




**Test d'hétéroscédasticité**

```
    studentized Breusch-Pagan test

data:  model1
BP = 20.766, df = 8, p-value = 0.007795
```

La p-value est inférieure au seuil de 5% alors on rejette l'hypothèse nulle d'homoscédasticité. Donc, on est

en présence d'un problème d'hétéroscédasticité des résidus. Il faut donc appliquer la correction de White.

**Correction de White**

```
t test of coefficients:

                          Estimate  Std. Error t value  Pr(>|t|)
(Intercept)             5.5206e-03  2.9537e-03  1.8691 0.0649777 .
Mediane_revenu_dispo   -1.5250e-07  1.0004e-07 -1.5244 0.1310443
Homme_sans_diplome     -5.2270e-08  4.4386e-08 -1.1776 0.2421598
Femme_sans_diplome      9.7135e-09  4.3406e-08  0.2238 0.8234497
Taux_chomage           -4.3372e-04  1.6051e-04 -2.7022 0.0082813 **
Taux_pauvrete           2.6370e-04  7.4535e-05  3.5379 0.0006497 ***
Taux_logements_sociaux  3.5063e-05  2.5560e-05  1.3718 0.1736575
Sexe_politique1         4.9341e-04  4.8360e-04  1.0203 0.3104212
Geographie1             3.0822e-04  2.9211e-04  1.0551 0.2942932
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Après la mise en oeuvre de la méthode, on remarque que le niveau de significativité des coefficients a diminué car le modèle a été corrigée de l'hétéroscédasticité. Toutefois, les variables significatives demeurent toujours significatives au seuil de 5%.

# Partie 2 : Analyse de la multicolinéarité

## Détection de la multicolinéarité

Afin de confimer nos soupçons concernant l'existence de multicolinéarité au sein de notre modèle final, nous regardons les facteurs d'inflation de la variance (VIF).

```
  Mediane_revenu_dispo     Homme_sans_diplome     Femme_sans_diplome
              2.516668              50.418393              47.280059
         Taux_chomage          Taux_pauvrete Taux_logements_sociaux
              2.759233               4.116913               2.568499
       Sexe_politique             Geographie
              1.124385               1.782771
```

Le **VIF** calcule la colinéarité d'une variable en fonction des autres régresseurs. La valeur au-dessus duquel nous considérons qu'il y a de la multicolinéarité n'est pas fixe, nous prendrons donc **5** comme valeur de reférence. On est donc en présence de multicolinéarité pour les variables `Homme` et `Femme` sans diplôme.

## Méthodes de réduction de dimension

## Partitionnement des données

Nous construirons notre modèle sur les données d'entraînement et évaluerons ses performances sur les données de test. Il s'agit d'une approche de *validation de hold-out* pour évaluer la performance du modèle. Notre échantillon d'apprentissage contient 70 % des données tandis que l'échantillon test contient les 30 % restants.

```
[1] 67  9
```

```
[1] 29  9
```

## Regression sur Composantes principales

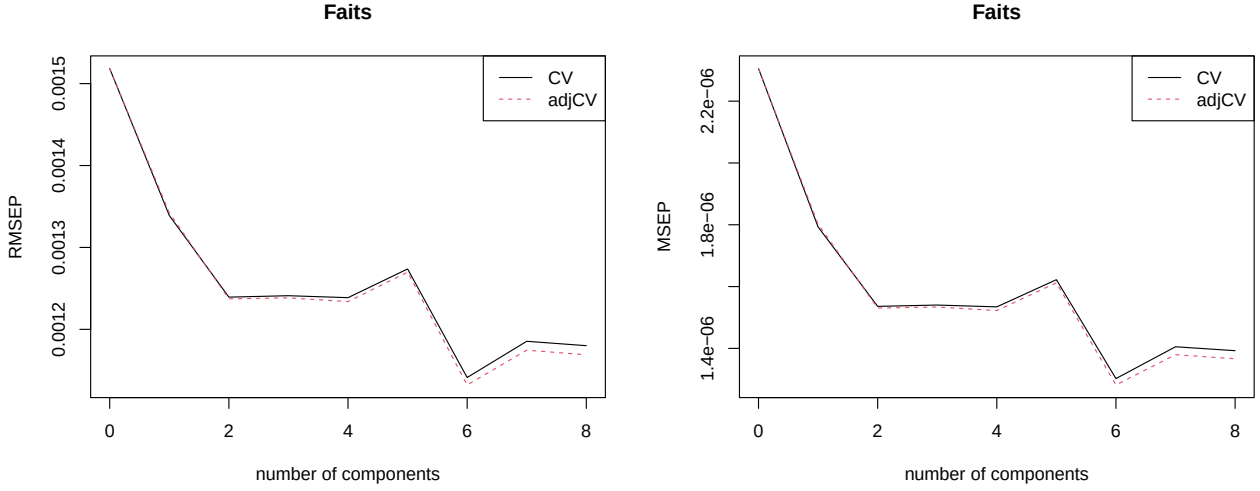
```
Data:   X dimension: 96 8
    Y dimension: 96 1
```

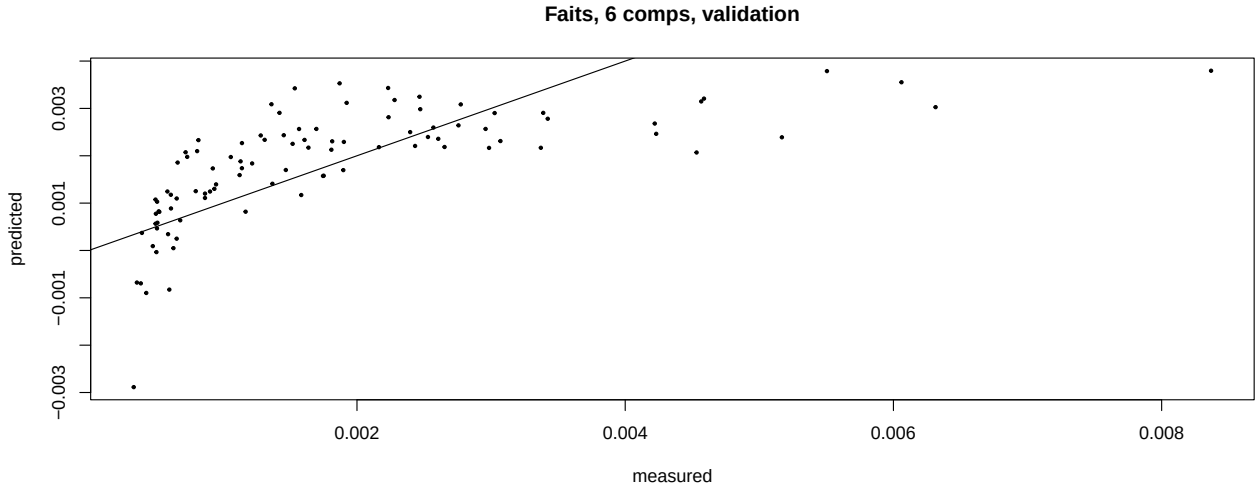
```
Fit method: svdpc
Number of components considered: 8

VALIDATION: RMSEP
Cross-validated using 10 random segments.
       (Intercept)   1 comps   2 comps   3 comps   4 comps   5 comps   6 comps
CV        0.001518  0.001339  0.001239  0.001241  0.001239  0.001274  0.001141
adjCV     0.001518  0.001342  0.001237  0.001238  0.001234  0.001270  0.001132
        7 comps   8 comps
CV     0.001185  0.001180
adjCV  0.001175  0.001169

TRAINING: % variance explained
       1 comps  2 comps  3 comps  4 comps  5 comps  6 comps  7 comps  8 comps
X        35.21    64.81    78.84    89.59    94.05    97.80    99.87   100.00
Faits    24.33    35.89    37.54    42.38    42.80    55.19    56.96    57.59
```

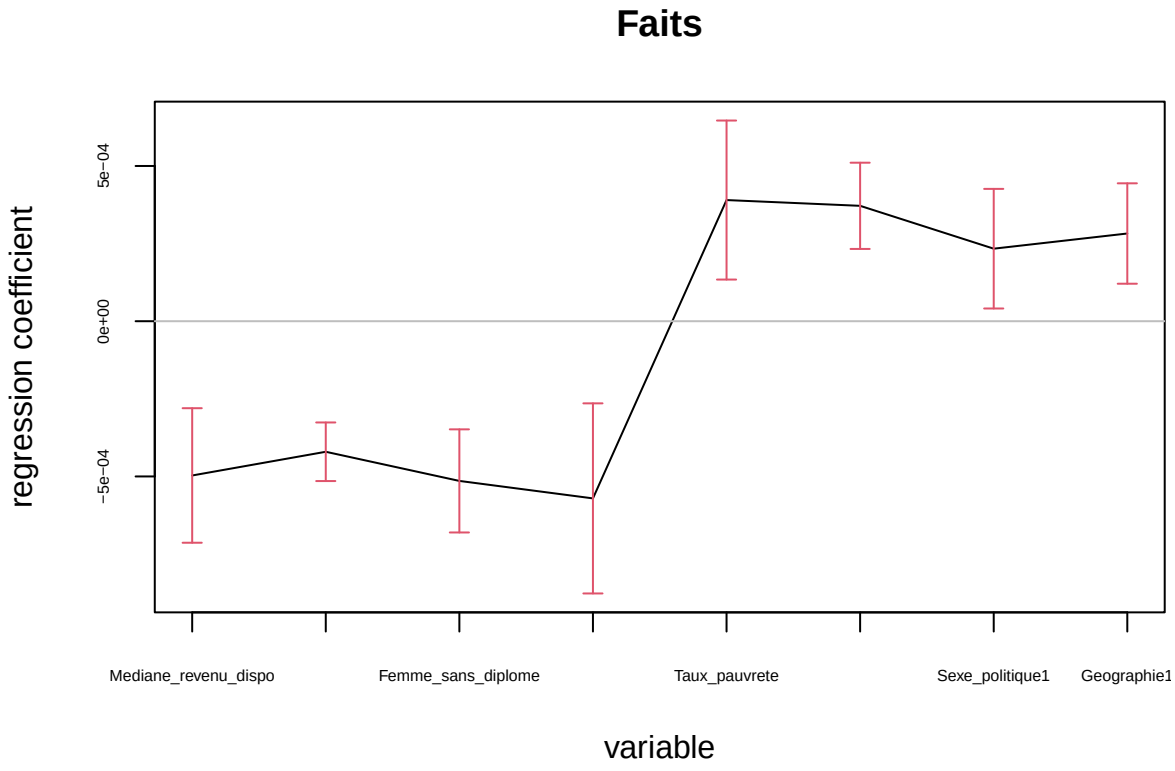
Après analyse des sorties du modèle PLS, on peut conclure que le modèle explique environ *57,6%* de la variance de la variable dépendante *Faits* en utilisant 8 composantes. Le modèle a été validé à l'aide de la RMSEP avec une validation croisée à 10 segments aléatoires et le modèle avec **7 composantes** présente la meilleure performance avec une valeur de 0.001075 pour CV et 0.001071 pour adjCV.




Les modèles avec plus de composantes peuvent potentiellement présenter une sur-ajustement, c'est-à-dire

qu'ils peuvent être trop complexes pour les données disponibles, ce qui peut entraîner des prédictions moins précises sur de nouveaux ensembles de données.




Les coefficients pour les variables *Mediane_revenu_dispo*, *Homme_sans_diplome*, *Femme_sans_diplome* et *Taux_chomage*, sont tous négatifs, ce qui suggère que des valeurs plus élevées de ces variables sont associées à des valeurs plus faibles du nombre de faits de viloences signalées. De même, le coefficient positif pour la variable *Taux_logements_sociaux* et *Taux_pauvrete* suggère une association positive avec la variable dépendante.

Toutefois, ces coefficients ne réfletent pas l'effet direct de chaque variable sur le nombre de faits de violences conjugales.

## Moindres carrés partiels

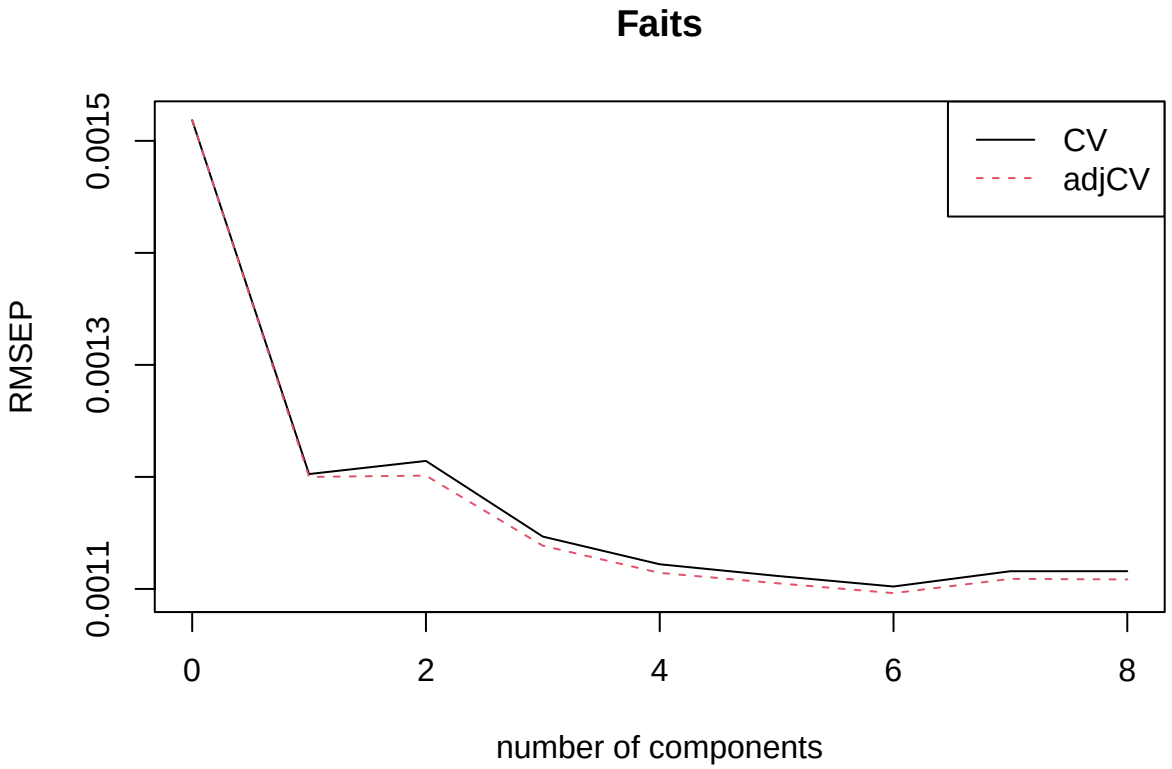
```
Data:   X dimension: 96 8
    Y dimension: 96 1
Fit method: kernelpls
Number of components considered: 8

VALIDATION: RMSEP
Cross-validated using 10 random segments.
       (Intercept)   1 comps   2 comps   3 comps   4 comps   5 comps   6 comps
CV        0.001518  0.001203  0.001214  0.001147  0.001122  0.001112  0.001102
adjCV     0.001518  0.001200  0.001201  0.001139  0.001114  0.001105  0.001096
       7 comps   8 comps
CV    0.001116  0.001116
adjCV 0.001109  0.001108

TRAINING: % variance explained
      1 comps  2 comps  3 comps  4 comps  5 comps  6 comps  7 comps  8 comps
X       33.19    49.65    71.13    80.77    92.50    95.52    98.97   100.00
```

```
Faits     42.52     51.11     54.63     56.77     56.93     57.11     57.24     57.59
```

On peut voir que l'erreur diminue au fur et à mesure que le nombre de composantes augmente, mais le taux de décroissance ralentit à partir de 5 ou 6 composantes. La meilleure performance est atteinte avec 6 composantes, où l'erreur de validation croisée est de 0.001156.

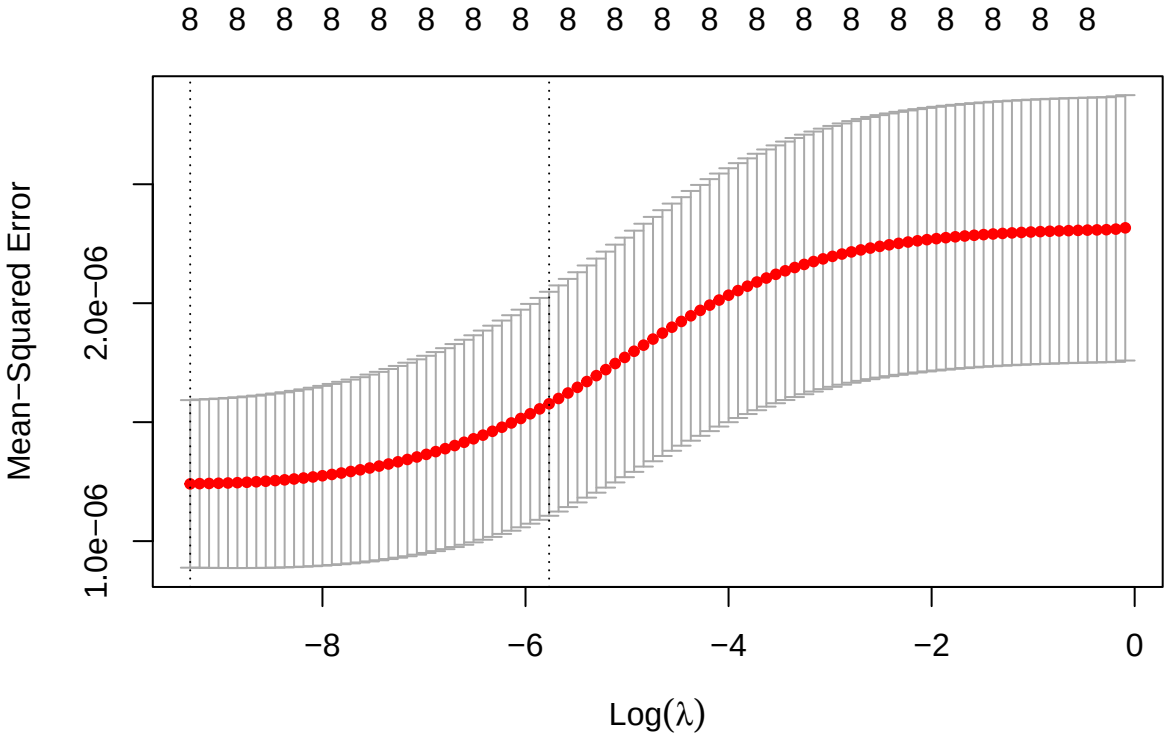



En ce qui concerne l'interprétation des coefficients, chaque coefficient représente uniquement la relation entre la variable explicative et celle de réponse. Par exemple, un coefficient négatif pour la variable *Homme_sans_diplome* indique une relation inverse entre cette variable et la variable de réponse, c'est-à-dire que les régions avec un taux plus élevé d'hommes sans diplôme ont tendance à avoir une valeur plus faible pour le nombre de faits de violences conjuguales signalées. De même, un coefficient positif pour la variable *Taux_pauvrete* indique une relation directe entre cette variable et la variable de réponse, c'est-à-dire que les régions avec un taux plus élevé de pauvreté ont tendance à avoir une valeur plus élevée pour le nombre de faits de violences conjuguales signalées.

## Méthodes pénalisées

Les méthode de pénalisation sont des techniques pour régulariser notre modèle linéaire et réduire le risque de surajustement (overfitting).

### Régression Ridge




```
[1] 9.116526e-05
```

On constate que la valeur du meilleur paramètre lambda qui minimise l'erreur quadratique moyenne estimée par validation croisée est **9.116526e-05**.

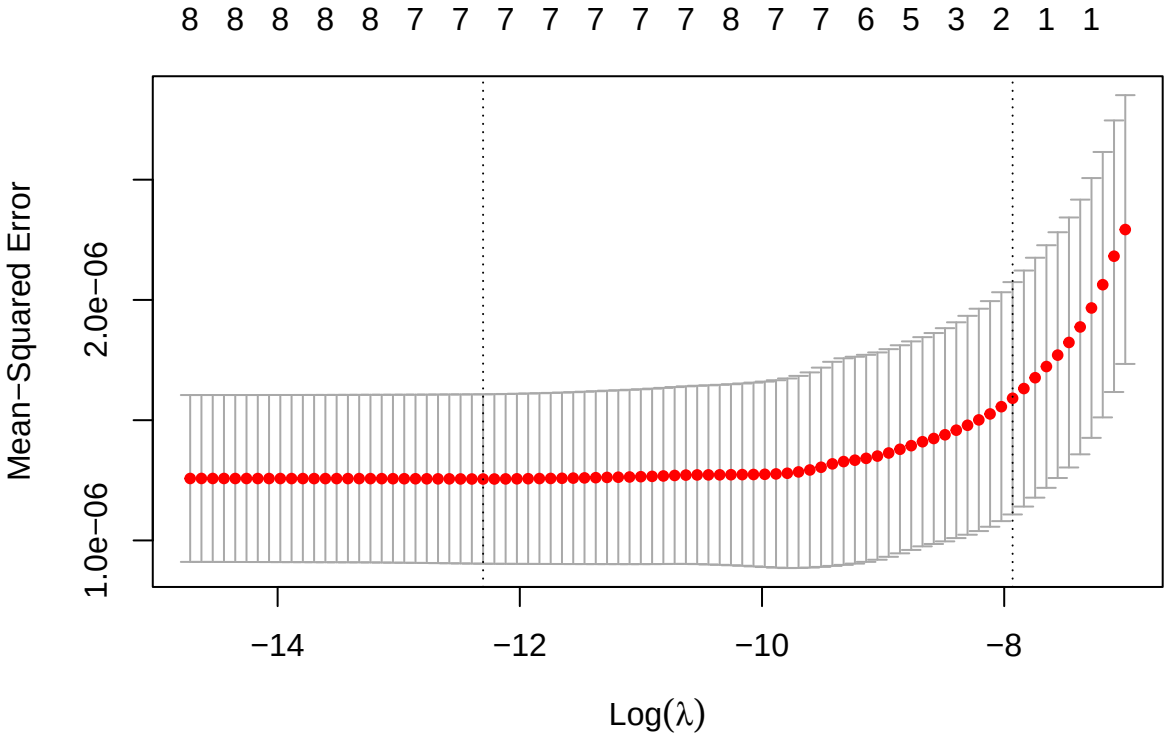
```
9 x 1 sparse Matrix of class "dgCMatrix"
                                s0
(Intercept)             5.445660e-03
Mediane_revenu_dispo   -1.744907e-07
Homme_sans_diplome     -2.172236e-08
Femme_sans_diplome     -1.398194e-08
Taux_chomage           -3.256978e-04
Taux_pauvrete           1.931377e-04
Taux_logements_sociaux  2.481363e-05
Sexe_politique          4.677787e-04
Geographie              2.651185e-04
```

On conclue ici qu'un coefficient négatif indique une *relation inverse* avec la variable dépendante, tandis qu'un coefficient positif indique une *relation directe* avec la variable dépendante. Plus la valeur absolue d'un coefficient est élevée, plus grande est l'importance de la variable correspondante pour la prédiction de la variable dépendante. On constate qu'en raison de l'utilisation de la régularisation dans le modèle de Ridge, certains de nos coefficients sont très petits (*Homme_sans_diplome* et *Femme_sans_diplome*) par rapport à d'autres. Cela est dû à la pénalisation de la magnitude des coefficients qui est utilisée pour éviter le **surajustement** et améliorer le modèle.

```
[1] 0.5660313
```

Le R-carré s'avère être de 0.5660313. C'est-à-dire que le meilleur modèle a été en mesure d'expliquer 56.6% de la variation des valeurs de réponse des données d'entraînement.

### Regression Lasso-Hitters




```
[1] 4.537317e-06
```

On constate que la valeur du paramètre lambda qui minimise l'erreur quadratique moyenne estimée par validation croisée est **4.537317e-06**.

```
9 x 1 sparse Matrix of class "dgCMatrix"
                                s0
(Intercept)             4.809058e-03
Mediane_revenu_dispo   -1.524774e-07
Homme_sans_diplome     -4.044207e-08
Femme_sans_diplome      .
Taux_chomage           -4.110808e-04
Taux_pauvrete           2.527320e-04
Taux_logements_sociaux  3.041160e-05
Sexe_politique          4.719454e-04
Geographie              2.741934e-04
```

On remarque qu'avec l'utilisation d'une régression de Lasso Hitters, certains de nos coefficients sont très petits (*Mediane_revenu_dispo* et *Homme_sans_diplome*) par rapport à d'autres. La variable *Femme_sans_diplome* a quant à elle été retirée du modèle. Encore une fois, cela est dû à la pénalisation de la magnitude des coefficients qui est utilisée pour éviter le **surajustement** et améliorer le modèle.

```
[1] 0.5750197
```

Le R-carré vaut 0.5750197. C'est-à-dire que le meilleur modèle a été en mesure d'expliquer 57.5% de la variation des valeurs de réponse des données d'entraînement.

Ainsi, la régression de Lasso Hitters apparaît être plus performante et proposer un meilleur modèle que la régression Ridge.

Table 2: Comparaion des différentes regressions

| Test | Rsquare | RMSE |
|---|---|---|
| Composantes principales | 56.96 | 0.0010 |
| Moindres carrés Partiels | 57.11 | 0.0011 |
| Ridge | 56.6 | 0.0097 |
| Lasso | 57.5 | 0.0096 |
| Elastic Net | 59.86 | 0.00078 |

### Régression Elastic Net

La régression Elastic Net combine deux formes de pénalisation, Ridge et Lasso.

```
  alpha lambda
1     0      0
```

On remarque que les meilleurs alpha et lambda estimés sur les données d'entrainement sont équivalents à 0.

```
9 x 1 sparse Matrix of class "dgCMatrix"
                                s1
(Intercept)             6.727882e-03
Mediane_revenu_dispo   -2.406051e-07
Homme_sans_diplome     -2.264493e-08
Femme_sans_diplome     -1.150290e-08
Taux_chomage           -3.768758e-04
Taux_pauvrete           1.576194e-04
Taux_logements_sociaux  5.071512e-05
Sexe_politique          8.349731e-04
Geographie              5.242352e-04
```

Avec l'application d'une régression Elastic Net, toutes les variables sont conservées dans le modèle final. Cependant, les variables *Mediane_revenu_dispo*, *Femme_sans_diplome* et *Homme_sans_diplome* ont des coefficients plus petits que les autres variables.

```
[1] 0.5985996
```

Le R-carré vaut 0.5985996. Donc le meilleur modèle a été en mesure d'expliquer 59.8% de la variation des valeurs de réponse des données d'entraînement.

### Choix du modèle

Afin de sélectionner le meilleur modèle nous étudions le tableau récapitulatif suivant :

Au vu des différents indiateurs, Elastic net apparaît être la meilleure régression car son R-squared est le plus grand et sa RSME est minimisée. Son R-squared est aussi supérieur à celui de notre modèle initial, ainsi, le modèle final conservé est celui produit par la régression Elastic Net.

## Partie 3 : Analyse de l'endogénéité

L'endogénéité survient lorsqu'une variable explicative (endogène) est corrélée avec les erreurs de régression. Cela peut entraîner des biais dans les estimations des paramètres du modèle et rendre difficiles l'interprétation et la validité des résultats obtenus. Il existe trois sources d'endogénéité :

- *les variables omises,*
- *les erreurs de mesure,*
- *la simultanéité*

Plusieurs approches peuvent être utilisées pour traiter l'endogénéité dans un modèle économétrique dont celle des variables instrumentales et des modèles à équations simultanées.

Les variables instrumentales sont des variables qui sont corrélées avec la variable endogène, mais qui ne sont pas corrélées avec les erreurs de régression. Elles sont utilisées pour estimer les paramètres du modèle en remplaçant la variable endogène problématique par ses valeurs prédites à partir des variables instrumentales. Cela permet de supprimer les biais potentiels dus à l'endogénéité. Avec les modèles à équations simultanées, les variables endogènes sont modélisées simultanément plutôt que séparément, ce qui permet de prendre en compte les interactions entre elles et de capturer les relations de causalité simultanée. Dans notre cas, les équations sont difficiles à identifier, la méthode des variables instrumentales serait donc plus adaptée afin de résoudre l'endogénéité qui pourrait être due aux variables omises.

Enfin, nous nous sommes questionnés sur les variables omises qui pourraient être ajoutées à notre modèle. Premièrement, les facteurs culturels ne sont pas présents dans nos variables. Une variable sur la religion aurait pu être pertinente et révéler des différences sur le nombre de faits, en fonction des proportions de chaque religion au sein des départements français. Deuxièmement, d'autres facteurs individuels seraient appropriés à notre étude tels que :

- la santé mentale des individus : les victimes de violences sexuelles peuvent avoir besoin d'un soutien psychologique pour faire face aux traumatismes, mais l'accès à des services de santé mentale peut varier en fonction du lieu de résidence, du niveau de revenu ou de l'assurance maladie)
- les antécédents de violences sexuelles : les personnes ayant été victimes de violences auparavant peuvent être plus susceptibles d'être victimes à nouveau. Cependant, tout comme le recensement des faits de violences sexuelles, les antécédents sont très difficiles à évaluer pour chaque département.

Nous avons réfléchi aux éventuels erreurs de mesure qui causées des biais dans notre estimation comme l'erreur de mesure des violences sexuelles. En effet, certaines personnes peuvent ne pas être conscientes qu'elles ont été victimes de violences sexuelles ou peuvent hésiter à signaler des cas de violences sexuelles.

# Conclusion

Pour conclure, dans cette étude sur les violences sexuelles, nous avons d'abord analysé notre base de données de façon univariée et bivariée afin d'orienter au mieux la suite de notre réflexion. Après-coup, nous avons sélectionné notre modèle économétrique, pour cela, nous avons testé différents modèles. Après avoir vérifier la spécification du modèle, nous avons retenu le meilleur modèle qui est celui ayant validé les test d'abscence d'autocorrélation, d'hétéroscédasticité. La présence d'une hétéroscédasticité nous a conduit à mettre en ouvre un modèle à correction d'erreur (MCE). Ce modèle présentait néanmoins des problèmes de multicolinéarité. Pour cela, nous avons appliqué des méthodes de réduction de dimension et de pénalisation. Parmi ceux ci, le modèle optimal est celui proposé par *elastic net*. Ce modèle est défini par :

$$Faits = 6.72e{-}03{-}2.4e{-}07{*}Mediane.revenu{-}2.26e{-}08{*}Homme.sans.diplome{-}1.15e{-}08{*}Femme.sans.diplome$$
$$-3.76e-04*Taux.chomage+1.57e-04*Taux.pauvrete+5.07e-05*Taux.logements.sociaux+$$
$$8.34e-04*Sexe.politique+5.24e-04*Geographie+e$$

Enfin, nous avons discuté des potentiels sources d'edogénéité qui pourraient exister dans notre modèle. La source prédominante d'endogénéité serait l'omission des variables `antécédents des violences familiales` et la `santé mentale`. Notre étude pourrait être poursuite en appliquant la méthode des variables instrumentales dans un modèle des Doubles Moindres Carrés (DMC).

# Bibliographie


**Données**

1. Nombre de faits signalés

2. Revenu médian disponible
3. Population
4. Homme sans diplôme
5. Femme sans diplôme
6. Taux de pauvreté, taux de chômage, taux de logements sociaux
7. Géographie
8. Sexe politique