Olivier CHEVALLIER
ochevall@gmail.com

STA201 - Rapport Projet

Etude du nombre de vol de voitures dans les villes américaines

STA201 2021/2022

INTRODUCTION

L'objectif de cette étude sera, à partir des données obtenues, de mettre en place un modèle de régression linéaire (en appliquant 2 méthodes étudiées dans le cadre de l'UE STA201) d'une des variables cibles (le nombre de vols de voiture pour des localités américaines pour 100 000 habitants en 1995).

L'objectif des modèles sera double: mesurer l'influence des différentes variables explicatives sur le taux de vol de voitures et la mise en place d'un modèle prédictif.

Pour cela nous réaliserons dans l'ordre les étapes suivantes :

- Nettoyage du fichier source des variables inutiles ou inexploitables (avec justification)

- Analyses (univariée, bivariée et multivariée) des données sources afin de détecter (et si possible corriger) dans les données sources ce qui pourrait altérer la qualité du modèle de régression (multicolinéarité, observations aberrantes, données manquantes)

- Mise en place d'un premier modèle de régression linéaire en utilisant la méthode PLS-R (PLS1, une seule variable cible), puis analyse du modèle obtenu

- Mise en place d'un second modèle, cette fois ci en utilisant la méthode LASSO

Le projet sera réalisé avec le logiciel R. Toutes les manipulations effectuées dans cette étude sont disponibles dans le fichier sta201.R joint à ce rapport.

METHODOLOGIE DU PROJET

RISQUE

Le risque de première espèce défini pour ce projet est de 5%.

MESURE DE LA PERFORMANCE DES MODÈLES

Les performances des modèles entrainés seront évaluées en utilisant l'erreur de prédiction (MSEP) ou sa racine carré (RMSEP), par validation croisée pour le choix de paramètres et en utilisant un échantillon de validation pour comparer les modèles obtenus par des méthodes différentes (PLS et LASSO).

ECHANTILLONS

Pour les besoins du projet l'échantillon initial sera découpé en 3 :

- L'échantillon d'entrainement (50%) sera utilisé pour entrainer les modèles
- Celui de validation (25%) pour mesurer la performance des modèles
- Celui de test (25%) pour publication des résultats du modèle choisi

BASES THEORIQUES DES METHODES UTILISEES

Rappel sur la régression linéaire

Le principe de la régression linéaire (simple ou multiple) est de modéliser au mieux la relation linéaire entre n variables explicatives et une variable cible y, en estimant l'équation représentant cette relation. Cette méthode s'applique pour des variables explicatives et expliquée quantitatives, Il est cependant possible d'utiliser des variables explicatives qualitatives en les modélisant sous forme d'indicatrices (en appliquant certaines précautions).

Son utilisation peut remplir 2 objectifs :
- Etudier la relation linéaire entre les variables explicatives et la variable cible
- Mettre en place un modèle de prédiction de la variable cible à partir des variables explicatives

On obtient une équation de la forme suivante, modélisant la relation linéaire entre X (Espace des n variables explicatives) et y :

$$y = ß0 + ß1X1 + ß2X2 + ... + ßnXn + Ɛ$$

où les ß0,ß1 ... ßn représentent les coefficients de la régression
Ɛ erreur aléatoire d'espérance égale à 0 et de variance σ^2

L'estimation de ces coefficients est obtenue par le critère des moindres carrés ordinaire (Ordinary Least Square), qui correspond à la minimisation de la somme des écarts quadratiques entre les valeurs observées et les valeurs prédites.

Hypothèses

La validité du modèle obtenu repose sur 4 hypothèses :

- Il existe une relation linéaire entre le(s) prédicteur(s) et la variable cible
- Les résidus obtenus sont indépendants
- La distribution des résidus suit une loi normale
- La variance des résidus est constante (homoscédasticité)

Limitations

- Multicolinéarité

Une forte multicolinéarité au sein de l'espace des prédicteurs engendre une instabilité des coefficients de régression. L'utilisation de méthodes de régularisations permet de résoudre cette problématique (Régression sur composantes principales, Régression PLS, Lasso, Ridge et Elastic Net). Dans ce projet nous étudierons les méthodes PLS et LASSO.

- Observations Aberrantes et influentes

La présence d'observations atypiques dans l'échantillon d'apprentissage d'un modèle de régression linéaire obtenu avec la méthode des moindres carrés peut avoir une forte influence sur les coefficients de régression obtenus, il est donc nécessaire de les détecter et de les prendre en compte, ce que nous verrons dans la partie consacrée à la régression robuste.

- Autres limitations

Non linéarité de la relation entre les prédicteurs et la variable cible
Auto-corrélation des résidus (séries temporelles, données spatiales)

Régression PLS (Partial Least Square)

La régression PLS-R est une méthode de régularisation, qui s'utilise dans les cas où la régression linéaire multiple ne peut être appliquée :

- Forte multicolinéarité des variables explicatives (instabilité des estimations des coefficients de régression)
- Plus de variables que d'individus (« Fat Data »)

Comme pour la régression linéaire multiple, l'objectif est d'obtenir l'équation modélisant au mieux la relation linéaire entre X (Espace des n variables explicatives) et la variable cible y.

Elle peut être appliquée pour la régression d'une variable cible quantitative (méthode PLS1) où plusieurs (PLS2). Dans le cadre de ce projet nous nous concentrerons sur la méthode PLS1 vue en cours.

Il s'agit d'une combinaison des caractéristiques de l'analyse en composante principale et de la régression multiple, basée sur la décomposition des variables explicatives en composantes orthogonales.

Contrairement à la méthode de construction des composantes de l'ACP (et de la régression sur composantes principales), la construction des composantes PLS maximise le lien entre les prédicteurs et la (ou les) variable(s) cible(s) (et non pas uniquement la variance des prédicteurs).

L'algorithme de la régression PLS est itératif, et est composé par des étapes successives de construction de composante et de régression (sur la variable cible puis sur les résidus obtenus par la régression par la composante précédente).

Paramètres du modèle

Le seul paramètre du modèle est le nombre de composantes à utiliser pour la régression, il est généralement obtenu par validation croisée ou en utilisant l'échantillon de validation.

Avantages et inconvénients de la méthode

+ Diminution de la variabilité des estimateurs due à la multicolinéarité des variables explicatives
+ Permet de traiter les Fat Data (plus de variables que d'observations)
+ Inclus le traitement de données manquantes
+ Méthode de visualisation
+ Calculs assez simples (pas d'inversion de matrice)

- Introduction de biais
- Non équivariante par rapport au changement d'échelle

Régression LASSO

La régression LASSO (Least Absolute Shrinkage and Selection Operator) est également une méthode de régularisation permettant de réduire la variabilité des estimateurs, au prix de l'introduction de biais.

Régularisation l1

La pénalisation LASSO minimise la fonction ci-dessous (avec le paramètre lambda >= 0 à optimiser) :

$$\sum_{i=1}^{n}(y_i - \sum_{j} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

Soit la somme des carrés des résidus + lambda * pénalisation

Paramètre du modèle

Le paramètre Lambda correspond au niveau de régularisation :

Si lambda = 0 , aucune régularisation, équivalent à OLS
Si lambda = ∞, tous les coefficients sont à 0
Quand Lambda augmente, le modèle se simplifie et le biais augmente
Quand Lambda diminue, la variance des estimateurs augmente

Il est optimisé de façon stochastique, par validation croisée ou en utilisant l'échantillon de validation.

Avantages et inconvénients de la méthode

+ Diminution de la variabilité des estimateurs due à la multicolinéarité des variables explicatives
+ Permet de traiter les Fat Data (plus de variables que d'observations)
+ Méthode de visualisation
+ Interprétation du modèle plus simple, sélection parcimonieuse des variables

- Introduction de biais
- Non équivariante par rapport au changement d'échelle
- Consistance de la sélection de variable fortement corrélées

PRESENTATION DE LA BASE DE DONNEES

Sources des données

Le jeu de données étudié dans ce projet provient du Machine Learning repository de l'université de Californie à Irvine (fichier *CommViolPredUnnormalizedData.txt*).

Lien direct vers la page du jeu de données

Il s'agit de la combinaison de plusieurs sources de données (qui n'a pas été réalisée dans le cadre de ce projet, mais en amont de la mise à disposition des données sur le site de l'UCI) :

- Recensement de la population américaine de 1990
- Etude de la police américaine sur l'année 1990, « Law Enforcement Management and Administrative Statistics » (LEMAS)
- FBI Uniform Crime Reporting Program (UCR) de 1995

Individus et variables

Tel que disponible sur le site de l'UCI, le fichier source contient 2214 individus et 147 variables. Les individus représentent des localités des Etats-Unis (« communities »). Les 2214 individus présents dans le fichier source représentent un échantillon de l'ensemble des localités des Etats-Unis. Dans le cadre de cette étude nous supposerons que ces individus sont indépendants et identiquement distribués.

Source des 147 variables du fichier

- 1 à 4 : non-prédictives servant à identifier la localité (Nom, état, ...)
- 5 : Pré-échantillonage en 10 blocs du jeu de données
- 6 (population) à 103 (PctSameState85) : Recensement de 1990
- 104 (LemasSwornFT) à 120 (PolicAveOTWorked) : Enquête de la police US de 1990
- 121 (LandArea) à 123 (PctUsePubTrans) : Recensement de 1990
- 124 (PolicCars) à 129 (PolicBudgPerPop) : Enquête de la police US de 1990
- 130 (murders) à 147 (nonViolPerPop) : données de crime du FBI pour l'année 1995 (variables cibles, non utilisable pour la prédiction)

La description de chacune des variables est disponible en annexe (« ANN00 - Description des variables.txt »).

Variable cible

Les variables issues du Crime Reporting Program du FBI recensent les différents crimes commis aux Etats-Unis durant l'année 1995. Ce sont les variables que l'on cherche à prédire et expliquer à partir de données disponibles avant. Elles ne peuvent donc pas servir en tant que variables explicatives.

Dans le cadre de ce projet, on se limitera à une seule variable cible: le nombre de vols de voiture pour 100 000 habitants (autoTheftPerPop). Les autres variables issues des données du FBI de 1995 seront supprimées du jeu de données.

Retraitements sur les données sources

Création d'un identifiant unique pour chaque localité

En concatenant le nom de la localité (variable 1, communityname) et l'état (variable 2, state), on obtient un identifiant unique, qui sera utilisé pour identifier chaque observation (rowname du dataframe).

Données issues de l'enquête de la police américaine de 1990 (LEMAS)

Seules les données des département de police d'effectif de plus de 100 ont été collectées, plus un échantillon de départements de taille inférieure.

« the LEMAS survey was of the police departments with at least 100 officers, plus a random sample of smaller departments »

Pour les variables issues de cette l'enquête, les données ne sont pas disponibles pour 85% des individus. Il s'agit de données manquantes de type MAR, sauf pour la variable LemasSwornFT pour laquelle il s'agit du type MNAR).

Compte tenu du pourcentage extrêmement élevé de valeurs manquantes, ces variables ne seront pas utilisées dans le modèle de régression et seront supprimées du jeu de données.

Variables population, densité et LandArea

La densité de population étant égale au rapport entre la population et la superficie, une de ces 3 variables est redondante et sera supprimée du jeu de données. On supprimera ici la variable LandArea.

DESCRIPTION DES DONNEES

Dans cette partie du projet nous allons étudier en détail chacune des variables du projet dans le but de valider la pertinence d'un modèle prédictif, et d'identifier les éventuels pré-traitements nécessaires aux méthodes choisies.

Analyse Univariée - Variables explicatives

Du fait du nombre élevé de variables explicatives, l'analyse ne sera pas faite de manière individuelle. Le tableau regroupant les caractéristiques de ces variables est disponible en annexe dans le fichier « ANN01 - Analyse univariée X.txt ». Ce que l'on peut en extraire :

- L'échelle des variances est très variable
- De nombreuses distributions très éloignées de la distribution normale

Afin d'homogénéiser les variances des variables explicatives, elles seront réduites (en utilisant les écarts types de l'échantillon d'entraînement).

Analyse Univariée - Variable cible

MIN.	1ST QU.	MEDIAN	MEAN	3RD QU.	MAX.
7	157	303	474	590	4969

STATISTIQUES DESCRIPTIVES

ECART TYPE	504.4
VARIANCE	254406.9
ASYMÉTRIE	2.6
APLATISSEMENT	13.2

INDICATEURS DE DISPERSION ET DE FORME DE LA DISTRIBUTION

Fig.1 – Histogramme de la variable cible autoTheftPerPop

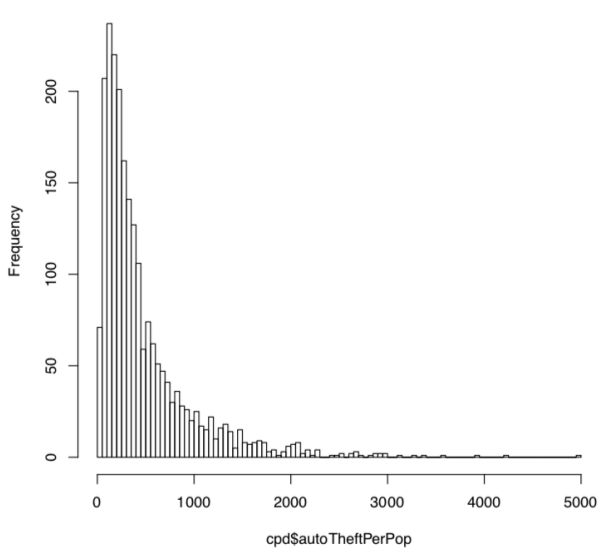
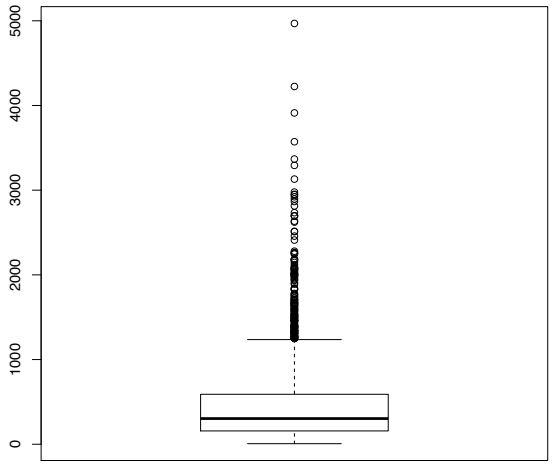


Fig.2 – Boxplot de la variable cible autoTheftPerPop



On constate :

- une distribution non-normale (asymétrie positive)
- 175 outliers (8%), au sens d'une distance supérieure à 1,5 IQR des quartiles inférieur et supérieur, ce qui ne nous étonne pas au vu de la distribution de la variable

Analyse Univariée - Log de la variable cible

En anticipation des impacts la non normalité de la distribution de la variable cible, étudions la transformée par la fonction LOG de cette même variable :

MIN.	1ST QU.	MEDIAN	MEAN	3RD QU.	MAX.
1.9	5.1	5.7	5.7	6.4	8.5

STATISTIQUES DESCRIPTIVES

	LOG(AUTOTHEFTPERPOP)
ECART TYPE	0.98
VARIANCE	0.96
ASYMÉTRIE	-0.19
APLATISSEMENT	3.20

INDICATEURS DE DISPERSION ET DE FORME DE LA DISTRIBUTION

Fig.3 – Histogramme du log de la variable cible autoTheftPerPop

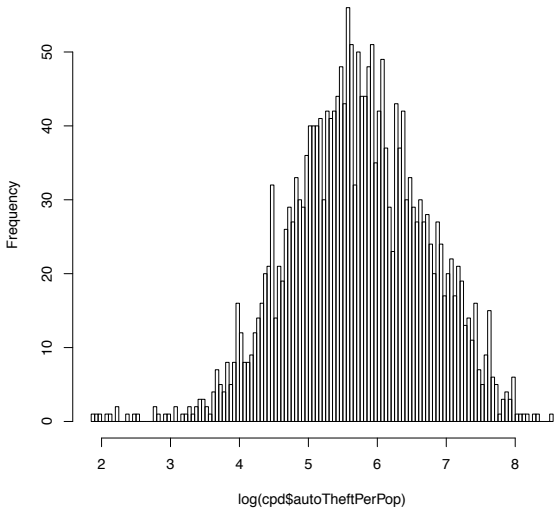
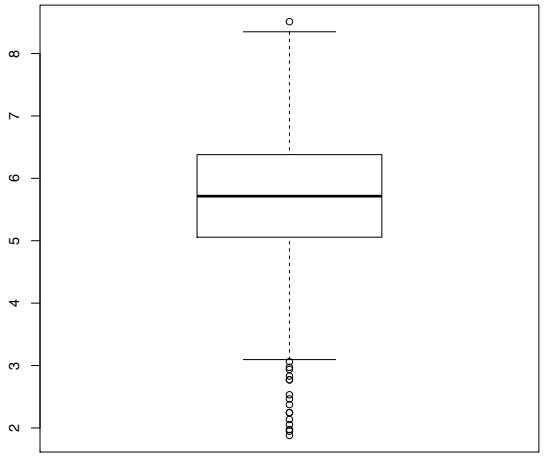


Fig.4 – Boxplot du log de la variable cible autoTheftPerPop



On constate ici :

- Beaucoup moins d'outliers sont détectés (17)
- La distribution par la fonction log de la variable cible se rapproche d'une distribution normale, information qui pourra s'avérer intéressante lors de l'étude des résidus de régression.

Analyse Bivariée - variables explicatives

Un export graphique de la matrice de corrélation linéaire des variables explicatives est disponible en annexe dans le fichier « ANN02 - Correlation_matrix_des_x.pdf ».

La recherche des corrélations linéaires supérieures à 0.95 en valeur absolue nous donne le résultat suivant :

```
MEDINCOME MEDFAMINC 0.979
PERCAPINC WHITEPERCAP 0.977
MALEPCTDIVORCE TOTALPCTDIV 0.975
FEMALEPCTDIV TOTALPCTDIV 0.983
PCTFAM2PAR PCTKIDS2PAR 0.985
PCTRECENTIMMIG PCTRECIMMIG5 0.989
```

```
PCTRECENTIMMIG PCTRECIMMIG8 0.978
PCTRECENTIMMIG PCTRECIMMIG10 0.965
PCTRECIMMIG5 PCTRECIMMIG8 0.993
PCTRECIMMIG5 PCTRECIMMIG10 0.984
PCTRECIMMIG8 PCTRECIMMIG10 0.995
PCTRECIMMIG10 PCTFOREIGNBORN 0.956
PCTLARGHOUSEFAM PCTLARGHOUSEOCCUP 0.985
PCTPERSOWNOCCUP PCTHOUSOWNOCC 0.982
OWNOCCMEDVAL OWNOCCHIQUART 0.984
RENTLOWQ RENTMEDIAN 0.963
RENTMEDIAN RENTHIGHQ 0.979
RENTMEDIAN MEDRENT 0.987
RENTHIGHQ MEDRENT 0.976
```

Pour chacun de ces couples de variables, l'une des deux sera écartée du projet.

Un export de la recherche des corrélations linéaires comprises entre 0.75 et 0.95 en valeur absolue est disponible en annexe dans le fichier « ANN03- Corrélation > 0.75.txt ».

On compte 82 relations (pour 85 variables restantes). Cependant il n'existe aucune justification pour écarter certaines de ces variables, nous noterons simplement la présence d'une forte colinéarité bivariée dans l'espace des prédicteurs.

Analyse Bivariée - variable explicative et cible

L'export des corrélation entre chacune des variables explicatives et la variable cible est disponible en annexe dans le fichier « ANN04 - Corrélations avec y.txt ».

Ci-dessous les corrélations les plus élevées avec la variable cible :

```
POPULATION  PEARSON: 0.322 SPEARMAN: 0.462
RACEPCTBLACK  PEARSON: 0.410 SPEARMAN: 0.441
RACEPCTWHITE  PEARSON: -0.557 SPEARMAN: -0.575
RACEPCTHISP  PEARSON: 0.348 SPEARMAN: 0.419
PCTWINVINC  PEARSON: -0.374 SPEARMAN: -0.414
PCTWPUBASST  PEARSON: 0.387 SPEARMAN: 0.395
PCTNOTHSGRAD  PEARSON: 0.327 SPEARMAN: 0.342
PCTUNEMPLOYED  PEARSON: 0.367 SPEARMAN: 0.366
TOTALPCTDIV  PEARSON: 0.400SPEARMAN: 0.496
PCTFAM2PAR  PEARSON: -0.464 SPEARMAN: -0.498
PCTYOUNGKIDS2PAR  PEARSON: -0.388SPEARMAN: -0.422
PCTTEEN2PAR  PEARSON: -0.455 SPEARMAN: -0.478
PCTKIDSBORNNEVERMAR  PEARSON: 0.525 SPEARMAN: 0.536
PCTRECENTIMMIG  PEARSON: 0.399 SPEARMAN: 0.362
PCTSPEAKENGLONLY  PEARSON: -0.392 SPEARMAN: -0.358
PCTNOTSPEAKENGLWELL  PEARSON: 0.413 SPEARMAN: 0.438
PCTLARGHOUSEFAM  PEARSON: 0.385 SPEARMAN: 0.404
PCTPERSOWNOCCUP  PEARSON: -0.402 SPEARMAN: -0.445
PCTPERSDENSEHOUS  PEARSON: 0.442 SPEARMAN: 0.556
PCTHOUSLESS3BR  PEARSON: 0.424 SPEARMAN: 0.469
MEDNUMBR  PEARSON: -0.353 SPEARMAN: -0.382
PCTVACANTBOARDED  PEARSON: 0.363 SPEARMAN: 0.311
PCTFOREIGNBORN  PEARSON: 0.435 SPEARMAN: 0.338
POPDENS  PEARSON: 0.425 SPEARMAN: 0.398
PCTUSEPUBTRANS  PEARSON: 0.352 SPEARMAN: 0.275
```

Les corrélations les plus élevées se situent autour de 0,5, et certaines variables ne sont pas du tout corrélées à la variable cible.

L'objectif de la suite, l'analyse multivariée, sera de détecter les situations pouvant poser problème lors de la mise en place de notre modèle de régression: confirmer (ou infirmer) la présence de multicolinéarité dans l'espace des prédicteurs et la détection des valeurs aberrantes multi-dimensionnelles.

Analyse Multivariée - VIF

Afin de détecter la présence de multicolinéarité au sein des variables explicatives, nous réaliserons une régression linéaire "test" afin d'étudier les valeurs du VIF les plus élevées :

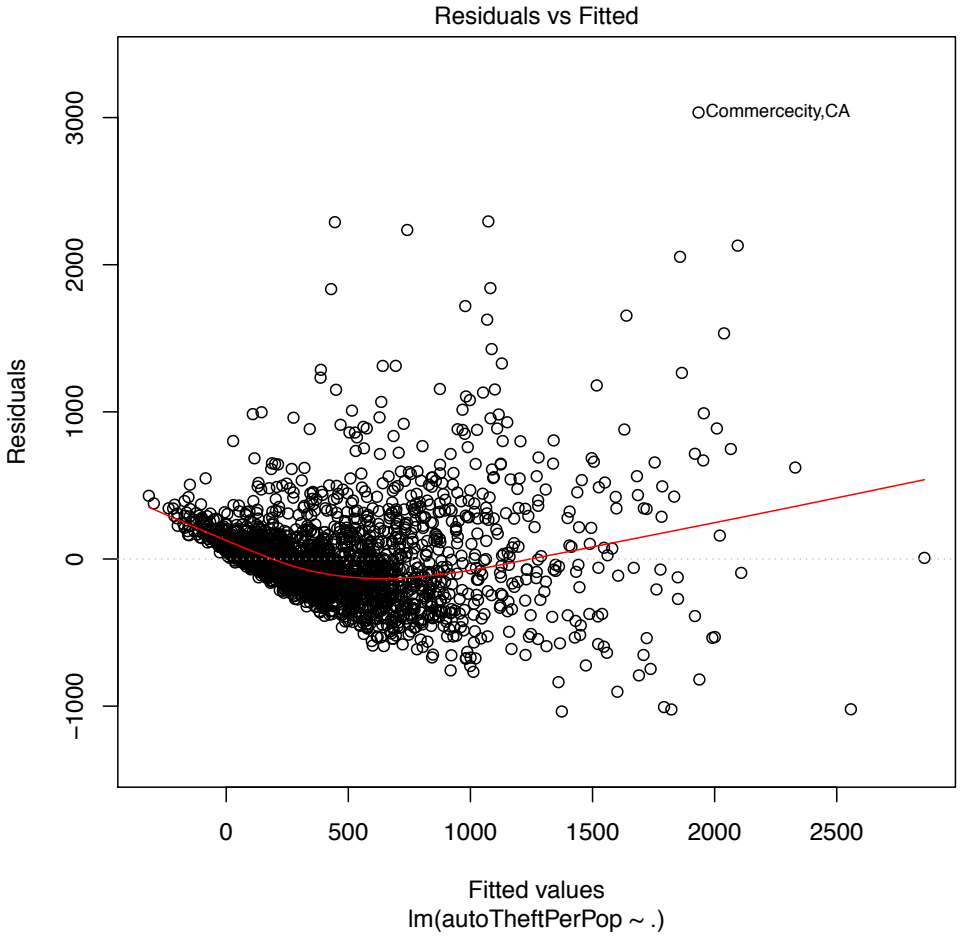
```
AGEPCT16T24 PERSPEROCCUPHOUS      AGEPCT12T29       PERSPERFAM        MEDINCOME       PCTFAM2PAR
        178              138               89               80               54               53
                                RÉGRESSION DE LA VARIABLE CIBLE - VIF
```

Avec 2 VIF supérieurs à 100 et 4 entre 50 et 100, on peut conclure à la présence d'une forte multicolinéarité au sein de l'espace des prédicteurs. Il sera nécessaire d'utiliser des méthodes de régularisation lors de la mise en place du modèle prédictif.

Analyse Multivariée - Résidus

Toujours dans le cadre de la régression linéaire "test", nous allons maintenant regarder la forme du nuage des résidus afin de vérifier l'hypothèse d'homoscédasticité des résidus.

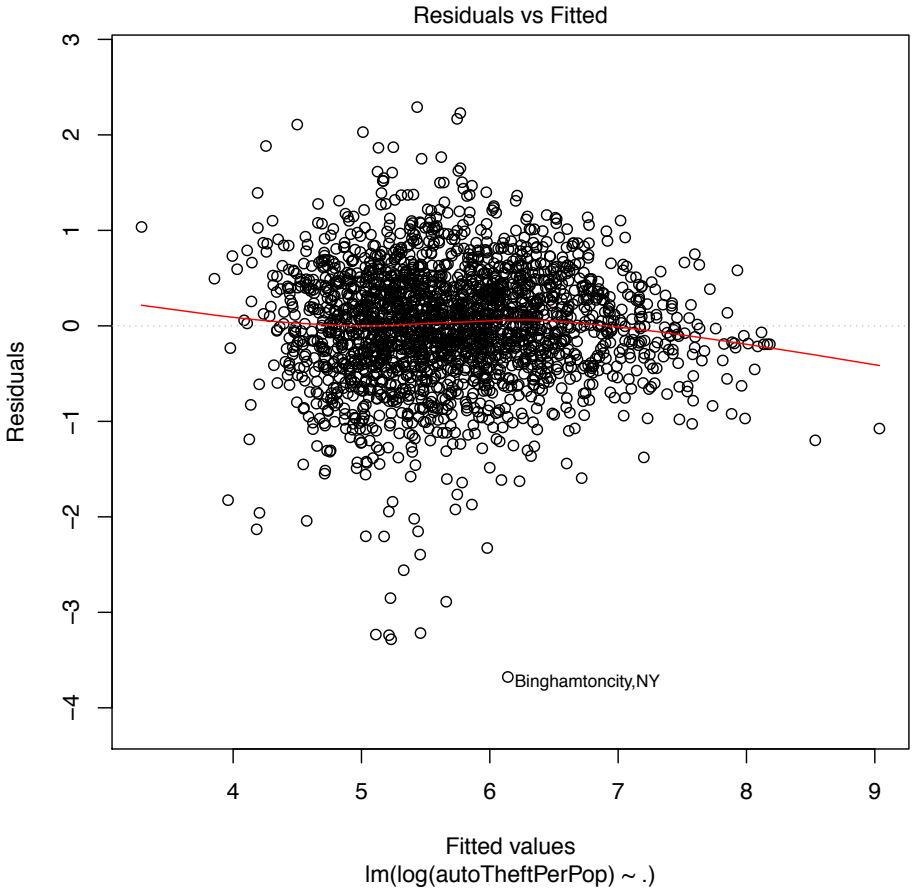
Fig.5 – Résidus de la regression de la variable cible



On peut constater sur ce graphe une violation de l'hypothèse d'homoscédasticité des résidus (nuage en forme d'entonnoir).

Etudions maintenant les résidus d'une régression "test" du log de la variable cible.

Fig.6 – Résidus de la regression du log de la variable cible



Cette fois il est difficile d'affirmer que l'hypothèse d'homoscédasticité des résidus n'est pas respectée.

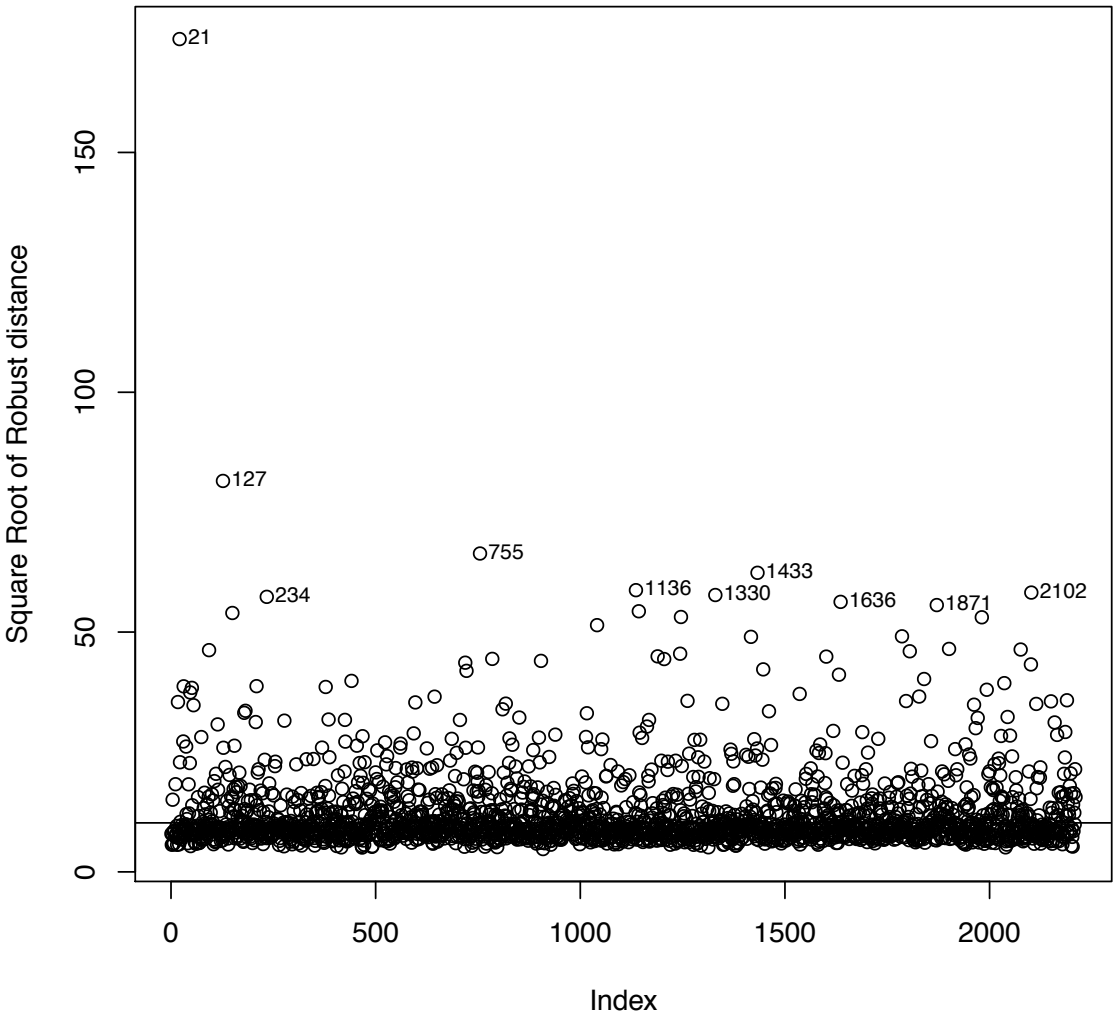
Pour la suite du projet et la mise en place des modèles prédictifs, la variable cible sera la transformée par la fonction log de la variable cible initiale.

Analyse Multivariée - Détection robuste des outliers

La présence d'observations atypiques dans l'échantillon d'apprentissage d'un modèle de régression linéaire obtenu avec la méthode des moindres carrés ordinaire peut avoir une forte influence sur les coefficients de régression obtenus, il est donc essentiel de les détecter et de les prendre en compte par des méthodes adaptées. Il est nécessaire d'utiliser des méthodes robustes de détection des valeurs aberrantes. Les mesures de distance « classiques » (distance de Mahalanobis par exemple) basées sur la moyenne et l'écart-type étant fortement influencées par ces mêmes valeurs aberrantes, certains points aberrants pourraient ne pas être détectés (« masking effect »).

Nous utiliserons dans le cadre de ce projet la distance de mahalanobis robuste obtenue au travers d'un estimateur MCD (Minimum Covariance Determinant) de paramètre $\alpha = 0.75$ (pour un compromis efficacité/robustesse).

L'estimateur détecte 862 outliers (sur 2210 observations) soit 40% de l'échantillon, en utilisant la règle du quantile 97,5% de la loi du Khi-deux à p degrés de liberté (avec p nombre de variables) comme valeur limite. Cette règle s'appuyant sur une hypothèse de distribution normale multidimensionnelle des variables (dont nous pouvons douter ici), nous utiliserons plutôt une analyse visuelle du graphe des distances de mahalanobis robustes afin d'identifier les observations aberrantes :



On constate qu'une observation se détache très nettement des autres (la ville de New York, observation 21) puis quelques observations éparses suivies d'un dégradé d'observations. La ligne horizontale correspond à la limite supérieure des observations non aberrantes selon la règle décrite précédemment.

Cela n'est pas surprenant, compte tenu des différences entre la ville de New-York et le reste des Etats-Unis. L'observation 21 sera supprimée du jeu de données et la ville de New-York exclue du périmètre du projet. Après étude au cas par cas, et sans justification pour les supprimer du jeu de données, les autres observations seront conservées.

Analyse Multivariée - ACP

Nous réalisons une analyse en composantes principales des variables explicatives afin de voir si des groupes d'observations apparaissent.

Compte tenu du nombre élevé de variables, il est difficile d'interpréter le cercle des variables. Le plan principale de cette ACP (contenant 39% de l'information initiale) est disponible en annexe (fichier « ANN06 - ACP plan principal.pdf »).

Il est difficilement lisible, mais parmi les individus bien projetés sur ce plan principal on peut constater la présence d'un groupe de villes de Californie dans le coin supérieur droit. On peut en déduire que ces villes se ressemblent au niveau de leurs caractéristiques et different du reste des villes américaines, mais il est difficile d'en tirer des conclusions plus détaillées.

RÉSULTATS DE L'ANALYSE AVEC LA RÉGRESSION PLS

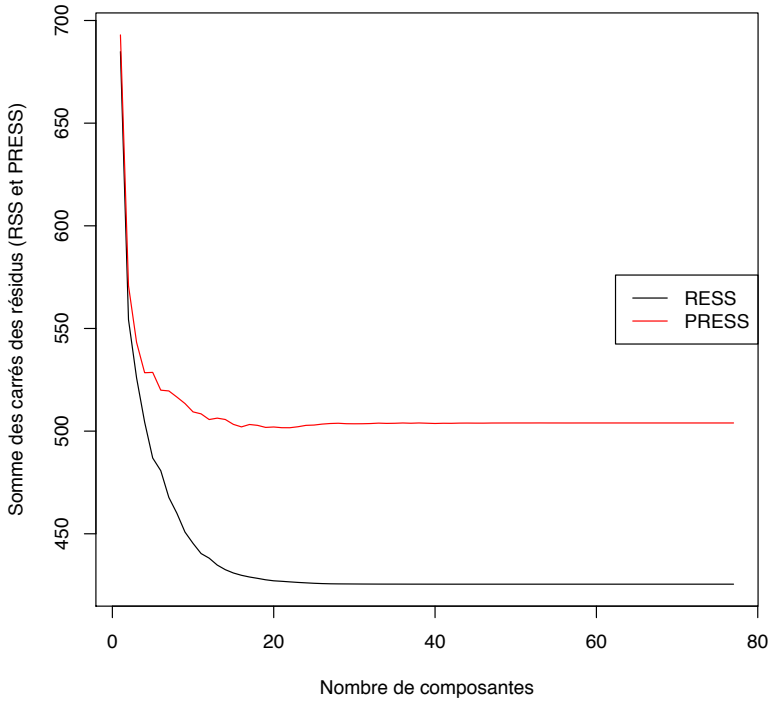
Entrainement du modèle et choix du nombre de composantes

Pour l'entrainement du modèle et obtention de la mesure MSE par validation croisée, nous utiliserons la fonction plsr du package R pls (les données explicatives sot déjà réduites) :

```
PLS.MDL <- PLSR(LOGAUTOTHEFTPERPOP ~ .,DATA=CPDR_TRAIN, VALIDATION="CV",SCALE=F)
```

Le critère utilisé pour le choix du meilleur modèle de régression PLS, c'est à dire le choix du nombre de composantes, sera la somme des carrés des résidus obtenu par validation croisée.

Fig.9 – Somme des carrés des résidus pour la régression PLS

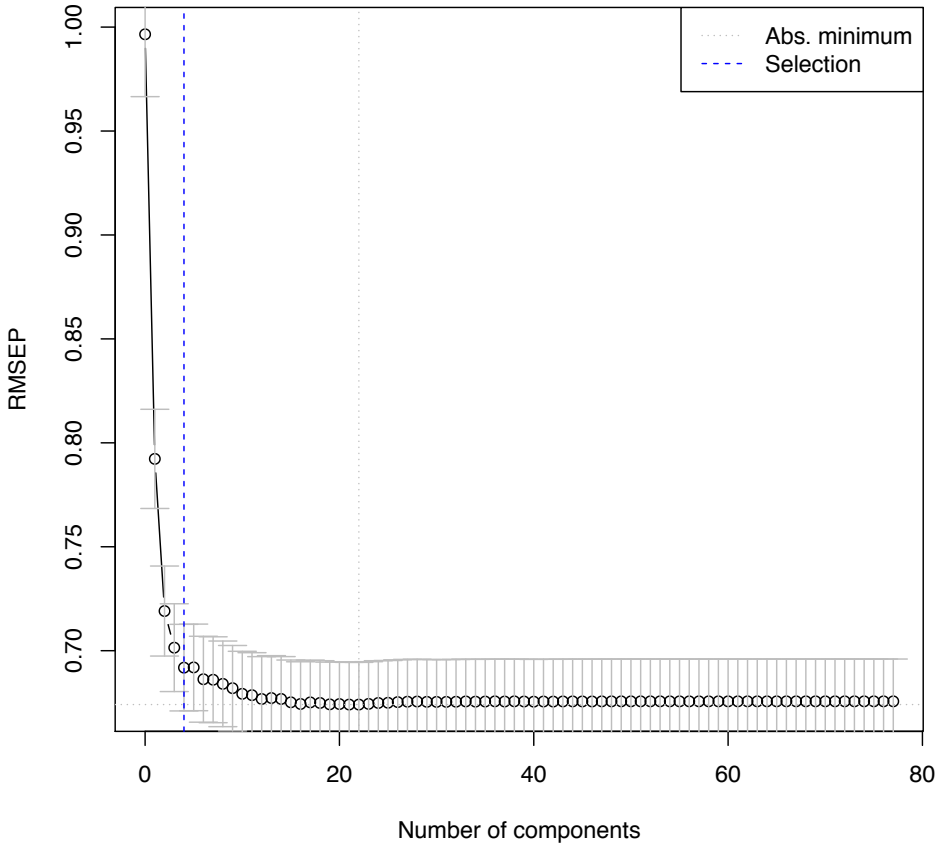


Ci-dessous le pourcentage de variance expliquée (sur l'échantillon d'entrainement) :

```
TRAINING: % VARIANCE EXPLAINED
                    1 COMPS  2 COMPS  3 COMPS  4 COMPS  5 COMPS  6 COMPS  7 COMPS  8 COMPS  9 COMPS  10 COMPS  11 COMPS  12 COMPS  13 COMPS  14 COMPS
X                     22.02    32.88    44.08    49.95    53.86    62.38    65.61    68.85    70.67     72.43     73.92     76.35     77.64     78.96
LOGAUTOTHEFTPERPOP    37.42    49.35    51.94    53.91    55.51    56.07    57.26    57.97    58.80     59.31     59.76     59.97     60.27     60.48
                    15 COMPS  16 COMPS  17 COMPS  18 COMPS  19 COMPS  20 COMPS  21 COMPS  22 COMPS  23 COMPS  24 COMPS  25 COMPS  26 COMPS  27 COMPS
X                     79.87     80.79     81.79     82.77     83.41     84.10     85.37     86.21     86.90     87.52     87.97     88.48     89.09
LOGAUTOTHEFTPERPOP    60.63     60.73     60.81     60.86     60.93     60.97     61.00     61.02     61.05     61.07     61.08     61.10     61.11
                    28 COMPS  29 COMPS  30 COMPS  31 COMPS  32 COMPS  33 COMPS  34 COMPS  35 COMPS  36 COMPS  37 COMPS  38 COMPS  39 COMPS  40 COMPS
X                     89.62     90.20     90.89     91.44     92.04     92.38     92.76     93.25     93.59     94.12     94.48     94.80     95.09
LOGAUTOTHEFTPERPOP    61.11     61.11     61.12     61.12     61.12     61.12     61.12     61.12     61.12     61.12     61.12     61.12     61.12
                    41 COMPS  42 COMPS  43 COMPS  44 COMPS  45 COMPS  46 COMPS  47 COMPS  48 COMPS  49 COMPS  50 COMPS  51 COMPS  52 COMPS  53 COMPS
X                     95.44     95.74     95.95     96.21     96.47     96.74     96.98     97.19     97.44     97.67     97.82     97.97     98.17
LOGAUTOTHEFTPERPOP    61.12     61.12     61.12     61.12     61.12     61.12     61.12     61.12     61.12     61.12     61.12     61.12     61.12
                    54 COMPS  55 COMPS  56 COMPS  57 COMPS  58 COMPS  59 COMPS  60 COMPS  61 COMPS  62 COMPS  63 COMPS  64 COMPS  65 COMPS  66 COMPS
X                     98.29     98.39     98.46     98.61     98.74     98.94     99.07     99.14     99.22     99.28     99.35     99.45     99.51
LOGAUTOTHEFTPERPOP    61.12     61.12     61.12     61.12     61.12     61.12     61.12     61.12     61.12     61.12     61.12     61.12     61.12
                    67 COMPS  68 COMPS  69 COMPS  70 COMPS  71 COMPS  72 COMPS  73 COMPS  74 COMPS  75 COMPS  76 COMPS  77 COMPS
X                     99.59     99.63     99.68     99.73     99.78     99.81     99.87     99.90     99.94     99.97    100.00
LOGAUTOTHEFTPERPOP    61.12     61.12     61.12     61.12     61.12     61.12     61.12     61.12     61.12     61.12     61.12
```

Le pourcentage maximum d'explication de la variance de la variable cible est obtenu dès la 30ème composante (62,12%), atteint 60% dès la 12ème et atteint 55% dès la 5ème.

Pour minimiser la complexité du modèle et limiter le sur-apprentissage, la méthode « one sigma » sera utilisée sur la métrique RMSEP obtenue par validation croisée.



Nous retenons donc le modèle à 4 composantes, qui explique 49.95% de la variance de X et 53.91% de la variance de y (voir tableau précédent).

Performances sur échantillon de validation

Afin de pouvoir comparer le modèle PLS obtenu au modèle LASSO du chapitre suivant, observons les performances de notre modèle à 4 composantes en prédiction sur l'échantillon de validation.

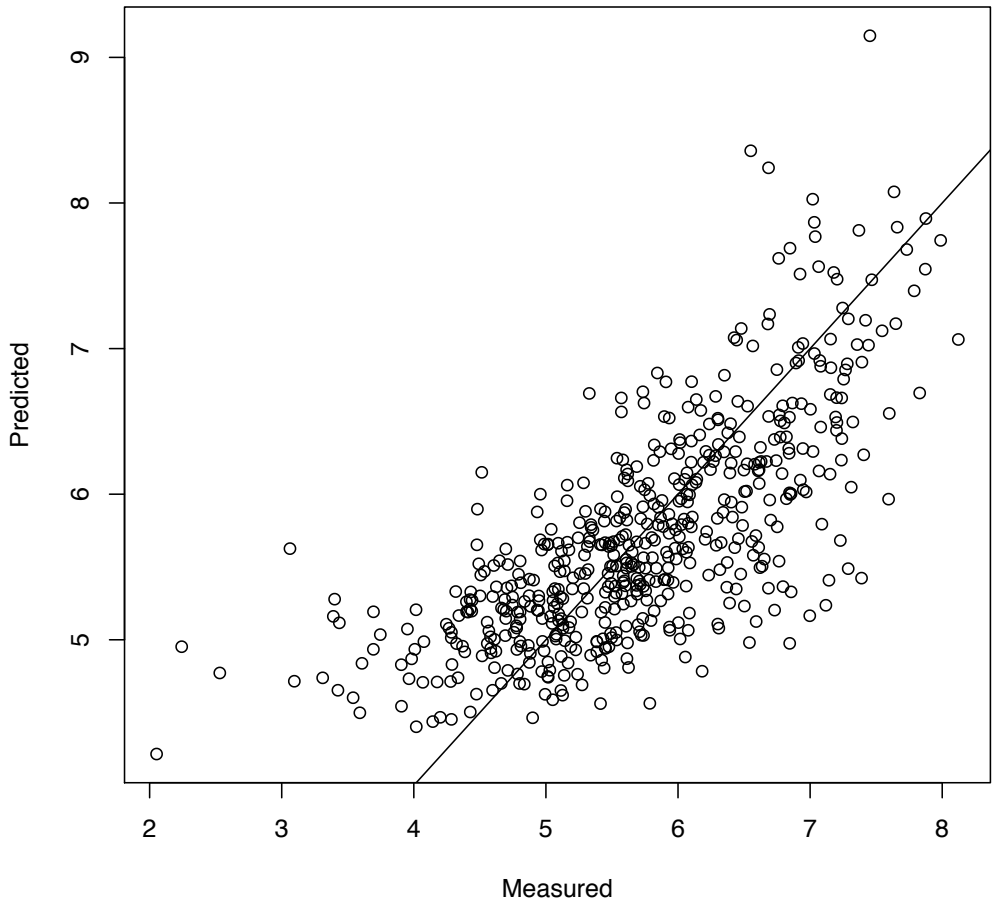
```
MSEP VALIDATION        0.46
RMSEP VALIDATION       0.68
```

La variable cible ayant été transformée par la fonction LOG avant utilisation, regardons également ces indicateurs calculés dans l'échelle originale de y (résidus calculés à partir de l'exponentielle de la variable mesurée et l'exponentielle de la valeur prédite) :

```
MSEP VALIDATION - ECHELLE CIBLE D'ORIGINE         271566
RMSEP VALIDATION - ECHELLE CIBLE D'ORIGINE        521
```

Ci-dessous le graphe valeurs mesurées (en abscisse) et valeurs prédites (ordonnée) pour l'échantillon de validation :

Fig.12 – PLS – Predicted vs Measured – Validation data



La RMSEP obtenue sur la variable re-transformée (521) est supérieure à l'écart-type de la variable cible de l'échantillon (504). La performance prédictive du modèle est faible.

Coefficients du modèle

En classant les coefficients sur données réduites, on obtient les variables les plus influentes sur la régression du nombre de vol de voiture pour 100 000 habitants.

Ci-dessous les top 10 des coefficients de régression obtenu avec le nombre de composantes sélectionnées plus haut (4). Ce tableau contient les coefficients bruts obtenus (colonne « For scaled data ») et les coefficients ramenés à l'échelle des X originale (colonne « For unscaled data»).

TOP 10 des coefficients positifs bruts

	For scaled data	For unscaled data
pctUrban	0.14093	0.003194348
TotalPctDiv	0.10663	0.035204736
population	0.06058	0.000000476
racepctblack	0.05968	0.004016416
PctVacantBoarded	0.05674	0.016576325
PctForeignBorn	0.05506	0.006767906
PctHousLess3BR	0.05221	0.003845849
PctKidsBornNeverMar	0.04052	0.012654521
PopDens	0.03637	0.000014008
PctSameCity85	0.03565	0.003261708

TOP 10 des coefficients négatifs bruts

	For scaled data	For unscaled data
PctBornSameState	-0.04267	-0.002629904
PctEmplProfServ	-0.04418	-0.006560178
MedNumBR	-0.04550	-0.088356850
PctBSorMore	-0.04660	-0.003681730
pctWInvInc	-0.04860	-0.003798611
agePct12t21	-0.05405	-0.011008680
PctFam2Par	-0.05572	-0.005227197
PctTeen2Par	-0.05586	-0.005303118
PctVacMore6Mos	-0.06646	-0.004803815
racePctWhite	-0.07226	-0.004365017

Pour la régularisation par régression PLS, aucune des variables n'est mise de côté, cependant les coefficients des variables dont l'influence est négligeable sont proche de 0.

L'ensemble des coefficients est disponible dans le fichier annexe « ANN05 - Coefficients PLS.txt ».

La formule du modèle sera donc de la forme :

$$y = e^{(7.60526 + 0.003194348 * population - 0.108754037 * householdsize + .. + 0.000014008 * PopDens + 0.005783830 * PctUsePubTrans - 0.001616655 * medHouseAge)}$$

RÉSULTATS DE L'ANALYSE AVEC LA RÉGRESSION LASSO

Entrainement du modèle et choix de lambda

Pour l'entrainement du modèle, nous utiliserons la fonction cv.glmnet du package R glmnet avec le paramètre alpha = 1 :

```
LASSO.MDL=CV.GLMNET(X_TRAIN,Y_TRAIN,
                    ALPHA=1,
                    LAMBDA=LAMBDAS,
                    NFOLDS = NFOLDS,TYPE.MEASURE="MSE")
```

Le paramètre lambda, après quelques essais entre 10^{-7} et 10^{6}, sera recherché entre 10^{-4} et 10^{3}.

La validation croisée (10 blocs) inclue dans l'entrainement du modèle nous donne le graphique suivant :

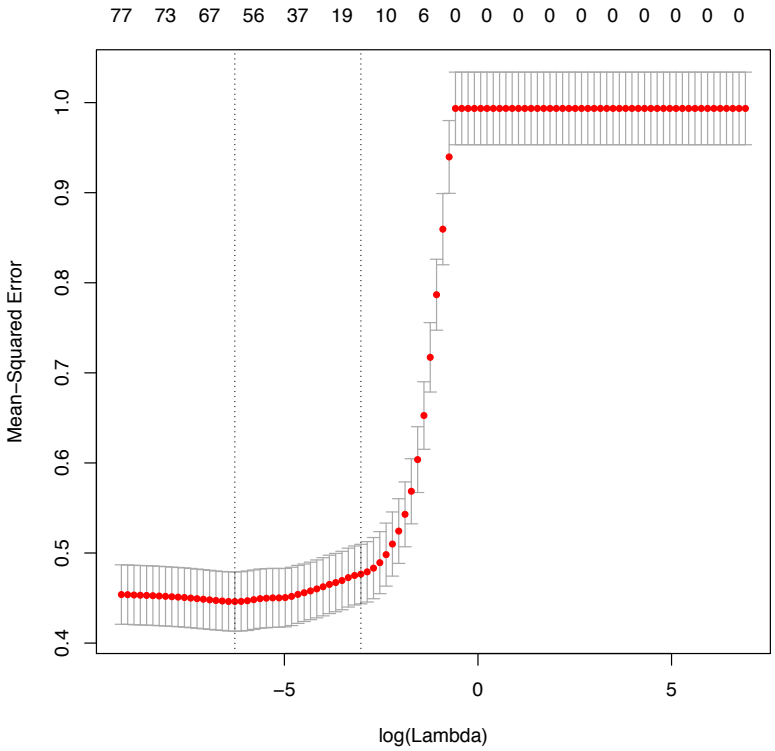


FIG14 - RÉGRESSION LASSO - MSE EN FONCTION DE LOG(LAMBDA) PAR VC

Afin de limiter le sur-apprentissage, nous sélectionnerons le paramètre lambda par la méthode de l'écart-type, en nous basant sur la MSE obtenue par validation croisée :

Par la méthode de l'écart « one-sigma », on trouve lambda = 0.049
Le modèle ayant obtenue le MSE le plus faible : lambda =0.0019, mais au prix d'un modèle plus complexe.

Performances sur échantillon de validation

Afin de pouvoir comparer les performances de notre modèle LASSO au modèle PLS obtenu au chapitre précédent, regardons ci-dessous les performances obtenues sur l'échantillon de validation pour le modèle LASSO de paramètre lambda = 0.049 :

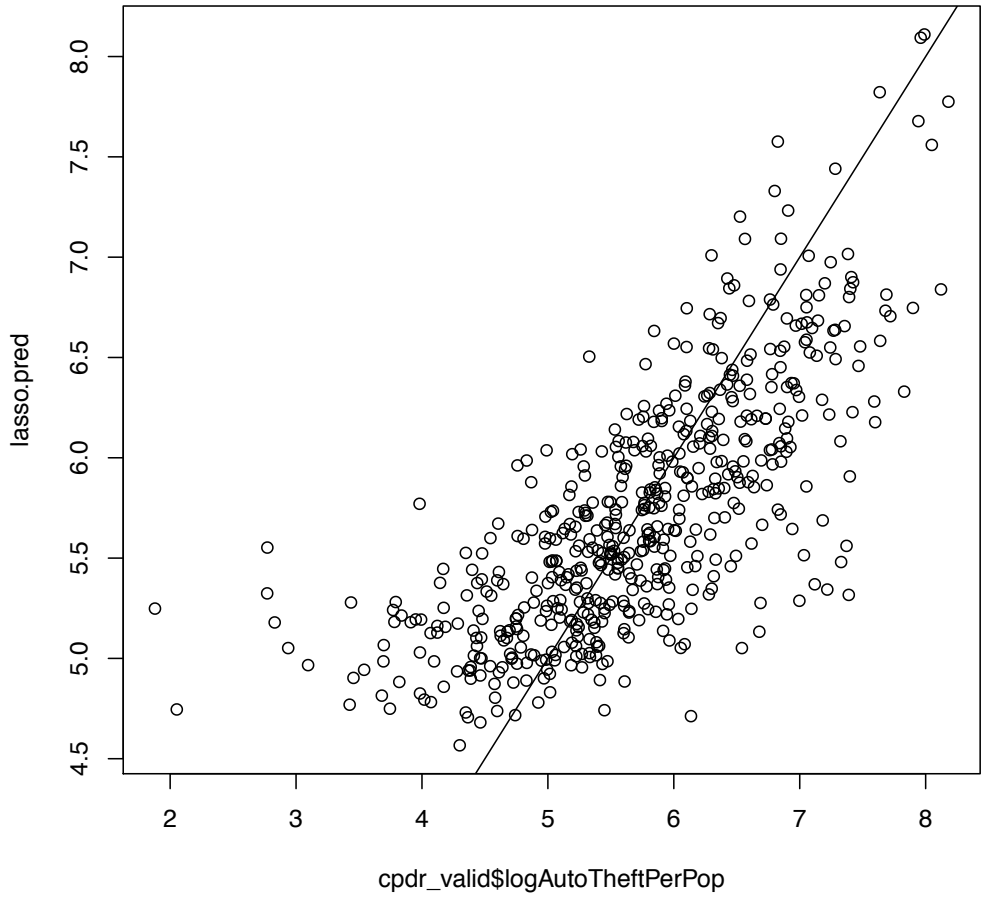
```
MSEP VALIDATION         0.47
RMSEP VALIDATION        0.68
```

Pour rappel, la variable cible a été transformée par la fonction LOG avant utilisation. Regardons également ces indicateurs après transformation dans l'échelle originale de Y :

```
MSEP VALIDATION - ECHELLE CIBLE D'ORIGINE         162842
RMSEP VALIDATION - ECHELLE CIBLE D'ORIGINE        404
```

Ci-dessous le graphe valeurs mesurées (en abscisse) et valeurs prédites (ordonnée) pour l'échantillon de validation :

Fig.15 – LASSO – Predicted vs Measured – Validation data



On peut constater que le modèle LASSO est plus précis que le modèle PLS sur les valeurs élevées de la variable cible, mais qu'il sur-estime les petites valeurs. Cependant avec une RMSEP de 404 la performance prédictive du modèle reste assez faible.

Coefficients du modèle

Ci-dessous les coefficients du modèle de régression obtenu avec lambda = 0,049. La colonne de gauche contient les coefficients à appliquer aux données réduites par l'écart-type de l'échantillon d'entrainement, la colonne de droite les coefficients à appliquer aux données sources brutes.

	SCALED.DATA.COEF	TRAIN.SD	RAW.DATA.COEF
(INTERCEPT)	5.4139	NA	5.41391782
POPULATION	0.0419	127175.04	0.00000033
RACEPCTWHITE	-0.1716	16.55	-0.01036729
PCTURBAN	0.1958	44.12	0.00443920
PCTBSORMORE	-0.0626	12.66	-0.00494486
TOTALPCTDIV	0.2525	3.03	0.08338116
PCTFAM2PAR	-0.0400	10.66	-0.00375527
PCTTEEN2PAR	-0.0150	10.53	-0.00142310
PCTHOUSLESS3BR	0.0714	13.57	0.00526141
MEDNUMBR	-0.0091	0.51	-0.01765865
PCTVACANTBOARDED	0.0310	3.42	0.00905114
PCTFOREIGNBORN	0.2308	8.13	0.02837126
POPDENS	0.0087	2596.56	0.00000335
PCTUSEPUBTRANS	0.0047	5.12	0.00091723

Tableau des coefficients de la régression LASSO

Pour les régressions LASSO, les coefficients non-significatifs pour le degré de régularisation sont à 0. On peut voir dans ce tableau les variables les plus importantes pour la régression du nombre de cambriolage en étudiant les coefficients pour les donnés réduites (colonne de gauche).

Positif : pctUrban, TotalPctDiv, PctForeignBorn
Négatif : racePctWhite

D'après le modèle, plus les pourcentages d'urbanisation, de divorce et de naissance à l'étranger sont élevés, plus il y a de vols de voitures. A l'opposé, plus le pourcentage de population d'origine caucasienne est élevé, moins il y a de vol de voitures. En regardant l'ensemble des coefficients obtenus on peut voir apparaitre l'influence de l'état du foyer (divorce, parents présents, nombre de chambres) et de l'environnement (immigration, densité, maisons abandonnées, urbanisation).

Ci-dessous la formule complète du modèle :

$Y = e^{(5.41391782 + 0.00000033 * population - 0.01036729 * racePctWhite + 0.00443920 * pctUrban - 0.00494486 * PctBSorMore + 0.08338116 * TotalPctDiv - 0.00375527 * PctFam2Par - 0.00142310 * PctTeen2Par + 0.00526141 * PctHousLess3BR - 0.01765865 * MedNumBR + 0.00905114 * PctVacantBoarded + 0.02837126 * PctForeignBorn + 0.00000335 * PopDens + 0.00091723 * PctUsePubTrans)}$

CONCLUSION

En comparant les performances des 2 modèles sur l'échantillon de validation, on constate que, s'ils font jeu égal sur la variable cible transformée (RMSEP = 0,68), la performance du modèle LASSO est meilleur sur le RMSEP de la variable cible re-transformée par la fonction exponentielle (404 contre 521). Cela est du au fait que le modèle LASSO est plus performant pour les valeurs élevées de la variable cible. C'est donc le modèle LASSO qui sera sélectionné, et dont voici les performances sur l'échantillon de test :

Modèle LASSO avec lambda = 0,049

MSEP TEST	**0.44**
RMSEP TEST	**0.66**
MSEP TEST - ECHELLE CIBLE D'ORIGINE	**141082**
RMSEP TEST - ECHELLE CIBLE D'ORIGINE	**376**

En comparant ce dernier résultat à l'écart-type de la variable cible sur l'ensemble du jeu de données (504), on peut voir que les performances prédictives du modèle obtenu sont à peine meilleures.

En creusant les valeurs prédites par les modèles et en les comparant aux données mesurées, on peut voir apparaitre la raison des faiblesses du modèle. Sans rentrer dans le détail, on peut voir 2 cas apparaitre pour les vols de voitures pour 100 000 habitants :

- sur-estimation dans les très grandes villes
- sous-estimation dans les villes autour de ces grandes villes

L'hypothèse d'indépendance des observations faite au début de ce projet semble incorrecte, et nos modèles prédictifs ne prennent pas en compte l'influence des villes voisines sur le nombre de vol de voiture.

Pour obtenir des modèles plus précis, il faudrait ajouter au jeu de données initial le géo-référencement des observations, et utiliser des méthodes de statistique spatiale pour prendre en compte l'auto-corrélation spatiale.