Barrio Pablo
Caron Cyriac

LDD3 EM

# Déterminants de l'élection de Barack Obama en 2008

Année 2020-2021

université PARIS-SACLAY

# Table des matières

# Introduction

En 2008, un jeune sénateur démocrate de l'Illinois, Barack Obama, n'ayant que dix ans d'expérience politique, remporte l'élection américaine face à John McCain, un vétéran du parti républicain, devenant ainsi le premier président de couleur des Etats-Unis d'Amérique. Dans un système électoral traditionaliste, ne comportant que deux partis ayant les moyens de leurs ambitions et un principe de grands électeurs ne représentant pas toujours très bien l'avis populaire, Obama se distingue en s'associant non pas à des slogans à visée politique, mais plutôt à visée émotionnelle tels que " Yes We Can " ou " HOPE ".

Ce revirement, par rapport aux stratégies classiques, s'explique notamment par le contexte économique et géopolitique de l'époque. En effet, en 2008, les Etats-Unis sont au cœur de la crise économique des "subprimes", tout en s'enlisant dans des guerres à la fois en Afghanistan et en Irak, depuis respectivement sept et cinq ans. C'est en réponse à des sentiments d'insécurité et de colère qu'Obama décide de se montrer non pas en tant que candidat démocrate, mais en tant que candidat de l'espoir.

A l'aide de nos connaissances statistiques et économétriques, nous nous sommes ainsi intéressés aux déterminants de cette élection singulière, et plus particulièrement à la véracité de l'idée selon laquelle Obama a réussi à s'affranchir des déterminants électoraux classiques, tels que la socio-démographie ou l'appartenance politique, pour se concentrer sur les sentiments et les opinions.

En se basant sur les données issues de l'enquête téléphonique de la National Annenberg Election Survey, une enquête universitaire de grande envergure réalisée avant les élections, notre travail s'est naturellement découpé en trois grandes parties :

- Nous avons tout d'abord trié, classifié et recodé les résultats à l'aide de statistiques descriptives et de régressions linéaires simples, afin d'obtenir une analyse à la fois économiquement et statistiquement significative.
- Puis, nous avons analysé les effets de ces variables dans une régression linéaire multiple, nous permettant de mieux comprendre les déterminants de l'élection.
- Enfin, nous avons procédé à une analyse de modèles par bloc en divisant nos variables en trois catégories : *Socio-démographie, Idéologie* et *Opinions*. Cette analyse repose sur des tests statistiques, et a pour objectif d'expliciter mathématiquement l'importance relative de chacun des blocs et donc leur rôle dans le résultat de l'élection.

# I. Choix des Variables

La table de données à laquelle nous avons eu accès, issue de l'enquête téléphonique de la National Annenberg Election Survey, est une base en coupe transversale avec 3737 observations et 66 variables (c'est-à-dire 3737 individus qui ont répondu à 66 questions de l'enquête).

Notre objectif étant d'expliquer comment Obama a réussi à convaincre les électeurs, nous avons choisi comme variable dépendante la variable RCa_10, qui correspond à la réponse à la question *" Pensez-vous voter pour Obama ? "*. Afin d'étudier cette variable, nous avons supposé que les individus ont suivi leurs croyances et ont voté pour Obama lorsqu'ils ont répondu oui. Cette variable est nominale et admet trois modalités différentes : *"Yes"*, *"No"*, et *"Don't know"*, ainsi que 842 valeurs manquantes. Dans l'enquête, 182 individus ont répondu "*Don't know*", un chiffre relativement petit par rapport aux individus ayant répondu "*Yes*"/"*No*" (2713). De plus, puisque notre objectif est de différencier les individus qui sont convaincus par Obama de ceux qui ne le sont pas, la réponse *"Don't know"* ne nous apporte pas plus d'informations que *"No"*. Nous avons donc décidé de transformer cette valeur en "*No*" afin d'obtenir une variable indicatrice, et d'estimer ainsi la probabilité de voter pour Obama en fonction de diverses caractéristiques, grâce à une régression linéaire multiple.

Pour sélectionner nos variables indépendantes, nous avons tout d'abord commencé par choisir de manière intuitive les variables qui pourraient avoir une corrélation avec notre variable dépendante, tel que le salaire de l'individu ou le parti auquel il s'identifie le plus.

Ensuite, nous avons regroupé les variables en trois blocs : le bloc *Socio-démographie* regroupant les caractéristiques éponymes, le bloc *Idéologie* représentant l'appartenance idéologique et le bloc *Opinion* regroupant les opinions des électeurs sur Obama et différents sujets politiques. Nous avons alors étudié de manière qualitative chaque variable en effectuant des régressions linéaires simples et des statistiques descriptives. Nous avons ainsi pu vérifier la significativité économique et statistique de chaque variable prise séparément. En effet, même si les estimations faites avec la régression linéaire simple risquent d'être biaisées par des corrélations avec d'autres variables, en vérifiant la significativité de chaque variable de manière individuelle, nous évitons la pollution de notre modèle final. Cette procédure nous a amené à :

- Supprimer les variables qui présentaient beaucoup de valeurs manquantes et/ou une significativité économique et statistique trop faible.
- Recoder les variables ordinales et nominales, selon les résultats de la régression et d'une analyse descriptive, afin de créer des modalités composites pertinentes.
- Transformer les nombreuses réponses indiquant que l'individu ne peut/veut pas répondre à la question en valeurs manquantes.

## Bloc Socio-démographie :

WA01_c (Sex) : variable indicatrice avec valeur de référence *"femme"*.

WA02_c (Age) : variable discrète que nous avons traité comme variable continue.

WA03_c (Education) : variable ordinale avec neuf catégories que l'on a regroupées en trois, afin d'augmenter la significativité statistique de la variable et de faciliter sa lecture.

WA04_c (Salary) : variable ordinale avec neuf catégories que l'on a regroupé en trois catégories (low, medium and high income) afin d'en augmenter la significativité statistique et de faciliter sa lecture. Cette variable présentant beaucoup de valeurs manquantes (1662), son ajout peut donc poser un problème de biais de sélection. Nous avons cependant considéré que cette variable est trop fondamentale pour être supprimée. Nous discuterons tout de même de sa pertinence et du possible biais ainsi amené.

WC03_c (Race) : variable nominale avec sept catégories. En partant de l'idée que nous souhaitions différencier les individus appartenant à une minorité ethnique des blancs, nous avons remarqué que les individus ayant répondu *"black, african american, black hispanic"* avaient une probabilité de voter pour Obama bien plus haute que le reste des minorités. Nous avons donc regroupé les modalités en trois catégories : *"black, african american, black hispanic"*, *"white or white hispanic"* et *"other"*.

## **Bloc Idéologie :**

RD01_c (Last-vote) : variable nominale représentant le vote de l'individu aux élections en 2004.

MA04_c (Conservative-liberal) : variable ordinale avec 5 catégories (*"very liberal"*, *"somewhat liberal"*, *"moderate"*, *"somewhat conservative"*, *"very conservative"*). Elle indique pour chaque individu s'il est plutôt conservateur ou libéral. Nous avons choisi de regrouper ces catégories en trois : "*liberal*", "*moderate*", "*conservative*", car les différences de probabilité de voter pour Obama entre "*very*" et "*somewhat*" pour les libéraux et pour les conservateurs sont négligeables.

MA01_c (Id-party): variable nominale qui indique le parti auquel le candidat s'identifie le plus. Elle peut prendre les valeurs suivantes ; "*republican*", "*democrat*", "*independent*", "*other*". Cette variable présente des paramètres estimés, économiquement et statistiquement significatifs.

## **Bloc Opinion :**

ABo12_c (Obama-share-values) : variable ordinale qui indique, sur une échelle de 0 à 10, si l'individu partage ses valeurs avec Obama. Nous avons d'abord décidé de la traiter comme une variable ordinale, et de traiter chaque niveau comme une variable indicatrice pour étudier la différence de probabilité estimée pour Obama entre chaque niveau.

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.010549   0.013612   0.775  0.43844
ABo12_c1    0.005076   0.029520   0.172  0.86347
ABo12_c2    0.003057   0.027977   0.109  0.91300
ABo12_c3    0.079004   0.028995   2.725  0.00647 **
ABo12_c4    0.062299   0.027693   2.250  0.02455 *
ABo12_c5    0.273542   0.022759  12.019  < 2e-16 ***
ABo12_c6    0.396637   0.026668  14.873  < 2e-16 ***
ABo12_c7    0.713731   0.023382  30.525  < 2e-16 ***
ABo12_c8    0.856459   0.020246  42.303  < 2e-16 ***
ABo12_c9    0.939806   0.022287  42.168  < 2e-16 ***
ABo12_c10   0.948766   0.019322  49.102  < 2e-16 ***
---
```

*Résultat de la régression linéaire de RCa10 sur ABo12_c*

Nous avons pu distinguer trois groupes : de 0 à 4, de 5 à 6, et de 7 à 10 que l'on a ainsi regroupé en : *"No"*, "*Somewhat*", "*Yes*". En effet, entre chacun de ces groupes, nous avons remarqué une différence de probabilité estimée de voter pour Obama bien supérieure que pour le reste des niveaux.

CDb01_c (Withdraw Iraq Troops): variable nominale avec trois avis possibles pour chaque individu: *"withdraw as soon as possible",* "*keep troops until stable government"*, *"set withdrawal deadline"*.

ABo05_c (Obama-leader) : variable ordinale qui indique, sur une échelle de 0 à 10, si l'individu pense que Obama est un bon leader. Nous avons traité cette variable comme une variable continue car la différence de probabilité estimée lorsque l'on augmente la valeur de la variable de 1 est relativement constante.

# II. Analyse du modèle complet

```
Coefficients:
                                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                                    0.9206871  0.0669110  13.760  < 2e-16 ***
WA01_cmale                                     0.0065611  0.0137307   0.478 0.632826
WA02_c                                         0.0005445  0.0005043   1.080 0.280431
WA03_cLot of education                        -0.0490653  0.0405363  -1.210 0.226304
WA03_cSome education                          -0.0552706  0.0395196  -1.399 0.162138
WA04_chi                                       0.0010193  0.0249882   0.041 0.967469
WA04_cmi                                       0.0081602  0.0234810   0.348 0.728244
WC03_cother                                   -0.1327113  0.0418582  -3.170 0.001551 **
WC03_cwhite                                   -0.0819830  0.0264429  -3.100 0.001967 **
RD01_cbush                                    -0.2197943  0.0226612  -9.699  < 2e-16 ***
RD01_cdid not vote for president              -0.0990125  0.0291126  -3.401 0.000688 ***
RD01_cother                                   -0.0082847  0.0480974  -0.172 0.863265
MA04_cconservative                            -0.1192706  0.0226271  -5.271 1.54e-07 ***
MA04_cmoderate                                -0.0963538  0.0184297  -5.228 1.94e-07 ***
MA01_cindependent                             -0.0828877  0.0190280  -4.356 1.41e-05 ***
MA01_cother                                   -0.1149487  0.0547405  -2.100 0.035896 *
MA01_crepublican                              -0.1390312  0.0242296  -5.738 1.14e-08 ***
CDb01_ckeep troops until stable government    -0.1099461  0.0221600  -4.961 7.74e-07 ***
CDb01_cset withdrawal deadline                -0.0697960  0.0179177  -3.895 0.000102 ***
ABo12_cNo                                     -0.3537413  0.0290392 -12.182  < 2e-16 ***
ABo12_csomewhat                               -0.3039167  0.0227712 -13.347  < 2e-16 ***
ABo05_c                                        0.0245109  0.0035504   6.904 7.29e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.266 on 1595 degrees of freedom
Multiple R-squared:  0.7204,    Adjusted R-squared:  0.7168
F-statistic: 195.7 on 21 and 1595 DF,  p-value: < 2.2e-16
```

*Résultat de notre régression linéaire effectuée sous R.*

## A. Vérification des hypothèses de Gauss-Markov

Le modèle reposant sur une enquête téléphonique, nous sommes bien conscients des limites de l'échantillonnage. En effet, il est beaucoup plus facile de raccrocher et donc de ne pas répondre à certaines questions que de quitter une enquête ayant lieu de vive voix. C'est pour cela que nous avons décidé d'effectuer un test sur la distribution conditionnelle des valeurs manquantes de la variable dépendante RCa10. Pour ce faire, nous avons créé une nouvelle variable RCa10_na qui est une indicatrice valant 1 si RCa10 présente une valeur manquante, 0 sinon. Puis, nous avons refait la régression complète de notre modèle en remplaçant RCa10 par RCa10_na.

Nos résultats sont ici très probants[1]. En effet quasiment tous les paramètres sont non statistiquement significatifs et les peu étant significatifs ont un impact économique proche de zéro. Étant donné que la significativité statistique représente la probabilité de rejeter l'hypothèse nulle à tort, et qu'ici, dans le cas où l'estimateur est tout de même significatif, sa valeur est très proche de zéro, nous pouvons affirmer que, prises ensembles, les variables n'impactent pas la probabilité d'avoir une valeur manquante. Ainsi, nous pouvons conclure que les valeurs manquantes de RCa10 ne sont pas distribuées en fonction d'une des autres variables de notre modèle, ce qui limite donc le biais induit par l'échantillonnage de cette enquête.

La régression linéaire de RCa10 sur notre modèle complet (i.e. celui contenant toutes nos variables) présente un R2 de 0.7204 et un R2 ajusté de 0.7168, ces deux valeurs étant très

[1] Le tableau de la régression étant conséquent, nous avons décidé de le mettre en annexe : Cf Annexe 1

proches, cela nous indique que notre modèle n'est pas mal spécifié. De plus, ces valeurs étant élevées, cela implique que notre modèle explique une grande part de la variance de RCa10. Ainsi, nous pouvons tirer des conclusions de nos paramètres estimés, tout du moins en fonction du sens et de la grandeur relative de leur impact. Nous ne pouvons pas garantir que ce modèle est non biaisé étant donné la complexité de nos données, mais au vu de la répartition aléatoire des valeurs manquantes de notre variable dépendante, du fait que l'espérance empirique de nos résidus est quasiment nulle ($E[\hat{u}] = 9*10^{-8}$) et de l'absence de colinéarité parfaite[2] dans nos variables, les résultats obtenus paraissent satisfaisants pour le théorème de Gauss-Markov, et notre étude.

## B. Analyse du modèle par paramètres

Nous allons à présent analyser les paramètres obtenus lors de la régression linéaire de RCa10 sur le modèle complet.

Groupe de référence : Les valeurs de référence prises par chaque variable sont : Sex = "*femme*", Education = "*no education*", Salary = "*low income*", Race = " *black*", Last vote = "*Kerry*", Conservative-liberal = "*liberal*", Id-party = "*democrat*", Withdraw-troops = "*withdraw troops as soon as possible*", Obama-share-values = "*share Obama values*". C'est à dire que l'intercept représente la probabilité moyenne estimée par le modèle pour une femme, à éducation inférieure au BAC, ayant de faibles revenus, noire, ayant voté pour Kerry aux dernières élections, se considérant libérale, s'identifiant au parti démocrate, qui veut retirer l'armée de l'Iraq au plus vite, qui partage les valeurs de Obama, qui pense que Obama n'est pas un bon leader et a 0 ans. Selon notre modèle, la probabilité estimée en moyenne pour une personne ayant ces caractéristiques de voter pour Obama est de 0.92.

Afin d'alléger les interprétations, nous précisons ici qu'à partir de maintenant nous raisonnerons entièrement de manière probabiliste et dans le cadre de notre modèle. C'est-à-dire que lorsque nous parlerons de probabilité, celle-ci sera une probabilité estimée en moyenne, donnée par notre modèle. De plus, toutes ces interprétations se feront par rapport à notre groupe de référence, à âge et avis sur le leadership d'Obama égales par ailleurs. Concernant les notations, "p.p" signifie points de pourcentage.

WA01_c (Sex) : Le paramètre valant 0.007, le fait d'être un homme augmente la probabilité de voter pour Obama de 0.7 p.p par rapport à une femme. Cette valeur très faible se remarque aussi dans la p-valeur qui est de 0.63, ce qui nous permet d'affirmer que le sexe n'a pas eu un effet important sur la probabilité de voter pour Obama.

WA02_c (Age) : Le paramètre valant 0.0005, chaque année de vie supplémentaire augmente de 0.05 p.p la probabilité de voter pour Obama. Cette valeur se retrouve ici aussi dans la p-valeur qui est de 0.28, ce qui nous permet d'affirmer que l'âge n'a pas non plus eu un fort effet sur l'élection. *(En considérant deux électeurs de 20 et 60 ans, leur différence d'âge n'implique qu'une hausse de 2 p.p de la probabilité de voter pour Obama, pour le plus âgé)*

WA03_c (Education) : Les paramètres pour les deux modalités valant -0.049 et -0.055, le fait d'avoir fait beaucoup (resp. un peu) d'études fait baisser la probabilité de voter pour Obama

[2] Cette vérification a été effectuée à l'aide la commande *"vif(reg)"* de R. : Cf Annexe 2

par rapport aux individus n'ayant pas fait d'études de 4.9 p.p (resp. 5.5 p.p). La p-valeur valant 0.23 (resp. 0.16), nous remarquons que celle-ci s'améliore par rapport aux précédentes, ce qui coïncident avec une valeur absolue des paramètres plus élevée. Cependant, le modèle ne nous fournit pas de preuve solide que l'éducation ait pu être déterminante.

WA04_c (Salary) : Les paramètres valant 0.001 et 0.008, le fait d'avoir un haut (resp. moyen) salaire augmente la probabilité de voter pour Obama par rapport aux individus ayant un faible salaire de 0.1 p.p (resp. 0.8 p.p). Les deux p-valeurs étant supérieures à 0.7, on peut affirmer que le salaire n'a pas eu d'impact significatif sur l'élection.

WC03_c (Race) : Les paramètres valant -0.08 et -0.13, le fait d'être blanc (resp. d'une autre ethnie) fait baisser la probabilité de voter pour Obama par rapport aux individus noirs/afro-américains de 8 p.p (resp. 13 p.p). Les deux p-valeurs étant inférieures à 0.003, nous pouvons considérer que l'ethnie a eu un réel impact. Enfin, le fait que l'effet de la modalité *"white"* soit inférieur à celui de *"other"* nous montre que les électeurs ne se sont pas uniquement concentrés sur sa couleur de peau.

RD01_c (Last vote) ; Les paramètres valant -0.22, -0.10, et -0.008, le fait d'avoir voté Bush[3], (resp. de ne pas avoir voté/d'avoir voté pour un autre candidat) fait baisser la probabilité de voter pour Obama par rapport aux individus ayant voté Kerry[4] de 22 p.p (resp. 10 p.p /0.09 p.p). Les p-valeurs des deux premières modalités sont très statistiquement significatives avec une p-valeur inférieure à 0.001 tandis que celle de la troisième ne l'est pas du tout avec une p-valeur de 0.86. On retrouve ici une continuité chez les électeurs qui ont tendance à voter pour des candidats affiliés au même parti. De plus, nous voyons aussi qu'Obama a aussi réussi à rattacher à sa cause les indécis comme le montrent les deux autres paramètres, la faible valeur du paramètre 3 nous montrant qu'au final les personnes ayant voté précédemment pour des candidats d'autres partis ont pu être touchées par Obama.

MA04_c (conservative-liberal) : Les paramètres valant -0.12 pour la modalité "*conservative*" et -0.097 pour "*moderate*". Les conservateurs (resp. modérés) ont donc une probabilité estimée de voter pour Obama inférieure aux libéraux de 12 p.p (resp. 9.7 p.p). De plus, les deux paramètres ont une p-valeur inférieure à 0.001, les paramètres sont donc très statistiquement significatifs. Ici, nous retrouvons le fait qu'Obama s'inscrit dans la pensée libérale américaine.

MA01_c (id-party) : Les paramètres ont pour valeurs -0.083 pour "*independent*", -0.115 pour "*other*" et -0.14 pour "*republican*". La probabilité de voter pour Obama pour les individus proches d'aucun parti est inférieure de 8.6 p.p (resp. 11.8 p.p pour les individus s'identifiant à un parti autre que le parti Démocrate et le parti Républicain, et 14.4 p.p pour les individus s'identifiant au parti Républicain) par rapport aux individus qui s'identifient au parti Démocrate. Les paramètres estimés de "*independent*" et "*republican*" sont statistiquement significatifs avec des p-valeurs inférieures à 0.001. Le paramètre estimé de "*other*" a une p-valeur comprise entre 0.01 et 0.001, il est donc statistiquement significatif au seuil de 1%.

ABo12_c (Obama-share-values) : Les paramètres valant -0.35 pour "*no*" et -0.30 pour "*somewhat*", le fait de ne pas partager les valeurs de Obama (resp. de les partager un peu) fait baisser la probabilité de voter pour Obama par rapport aux individus qui les partagent de 35 p.p (resp. 30 p.p). Les deux paramètres ont une p-valeur inférieure à 0.001, ils sont donc statistiquement significatifs. Les paramètres estimés nous présentent une forme de scission très

[3] Georges W. Bush, candidat républicain et vainqueur de l'élection présidentielle de 2004.
[4] John F. Kerry, candidat démocrate à l'élection présidentielle de 2004.

représentative de l'électorat américain. En effet, de ces valeurs nous pouvons dire que pour voter pour Obama, il faut absolument partager ses valeurs, sans place pour la modération. Ainsi, même avec son aspect fédérateur, Obama n'a tout de même pas pu s'affranchir de certains fondamentaux de l'électorat américain. Nous retrouvons ici aussi l'hypothèse selon laquelle les opinions ont été déterminantes lors de l'élection.

CDb01_c (Withdraw Iraq Troops) : Les paramètres valant -0.07 pour "*set withdrawal deadline*" et -0.11 pour "*keep troops until stable government*", le fait de vouloir fixer une date limite pour retirer les forces de l'armée, (resp. de vouloir laisser l'armée jusqu'à l'arrivée d'un gouvernement stable) fait baisser la probabilité de voter pour Obama par rapport aux individus qui veulent retirer l'armée immédiatement de 7.3 p.p (resp. 11.5 p.p). Les deux paramètres ont une p-valeur inférieure à 0.001. Ils sont donc statistiquement significatifs au seuil de 0.1%.

ABo05_c (Obama-leader) : Le paramètre vaut 0.025. Ainsi, augmenter la note selon laquelle Obama est un bon leader de 1 augmente la probabilité de voter pour Obama de 2.5 p.p. Le paramètre a une p-valeur inférieure à 0.001. Il est donc statistiquement significatif. Cet estimateur nous prouve ici l'importance de "l'aura" d'Obama. En effet, pour deux individus, un pour qui Obama n'est pas du tout un bon leader et le second pensant que si, la différence de probabilité de vote est de 25%. Nous retrouvons ainsi l'hypothèse de notre mémoire, selon laquelle les électeurs d'Obama ont surtout été touchés par leurs émotions et leurs opinions dans le choix d'élire Obama à la tête des Etats-Unis.

## C. Un possible biais de sélection

Nous avons choisi d'ajouter la variable WA04_c (salaire) dans la régression car nous trouvons que cette caractéristique socio-démographique est bien trop importante pour être négligée. Cependant, cette variable a exactement 1662 valeurs manquantes, et la différence de valeurs manquantes entre le modèle où elle est incluse et le modèle où elle ne l'est pas est de 973. Nous nous attendons donc à ce que l'inclusion de cette variable entraîne une perte de précision de nos estimations, et aussi un possible biais de sélection. En effet, la diminution d'individus dans le modèle entraînerait une hausse de la standard-error et donc une baisse du t-student. De plus, l'inclusion d'une nouvelle variable dans un modèle linéaire multiple risque aussi de diminuer la précision de nos paramètres estimés, à cause d'une corrélation avec les variables déjà présentes dans le modèle. Les p-valeurs devraient donc être supérieures à celle du modèle sans la variable salaire.

Néanmoins, nous avons pu observer que ceci n'est pas le cas. En comparant les p-valeurs, nous avons constaté que la plupart des p-valeurs de nos paramètres diminuent lorsque nous incluons le salaire. Nous ne pouvons donc pas affirmer qu'il y ait une perte de précision associée au salaire.

Ensuite, nous avons cherché à savoir si l'ajout de la variable du salaire pouvait créer un biais de sélection. Nous avons vérifié sur Internet la répartition des salaires aux Etats-Unis et avons trouvé que 20% de la population a un salaire inférieur à 35000 $, 40% a un salaire compris entre 35000 $ et 75000 $, et 40% a un salaire supérieur à 75000 $. Or pour ces mêmes salaires nous avons la répartition suivante dans notre échantillon : 14.4% ( $\frac{298}{2075}$), 45% ($\frac{934}{2075}$), 40.6% ($\frac{843}{2075}$). Notre échantillon semble donc représentatif. Pour vérifier ceci, nous avons fait une régression sans le salaire sur l'ensemble de l'échantillon et une régression sans le salaire sur l'échantillon où ne figurent pas les individus ayant des valeurs manquantes pour la variable du

salaire. Ainsi, nous avons calculé un intervalle de confiance au seuil de 95% pour chacun des paramètres de la première régression, et nous avons vérifié si les paramètres de la seconde régression étaient bien dans ces intervalles (Cf Annexe 3). Nous avons pu vérifier que les paramètres estimés par la seconde régression appartiennent tous aux intervalles de confiance respectifs. Nous pouvons donc affirmer que, s'il existe un biais de sélection sur notre régression, celui-ci est faible et nous pouvons le négliger.

Nous ne pouvons donc pas affirmer que l'inclusion du salaire dans notre modèle et la réduction de la taille de l'échantillon entraînent une perte de précision, ni un biais de sélection. Le salaire étant une caractéristique socio-démographique très importante, nous avons donc décidé de garder cette variable qui, avec le reste des variables du bloc *Socio-démographie,* va avoir un effet de contrôle sur le reste de nos estimations.

## Conclusion de l'analyse du modèle complet

Grâce à la vérification des hypothèses de Gauss-Markov, nous savons que les estimateurs obtenus grâce à la méthode des moindres carrés contiennent une part suffisante d'informations pour que l'on puisse en inférer des caractéristiques réelles.

Nous pouvons ainsi remarquer que les caractéristiques de l'individu le plus susceptible de voter pour Obama sont celles de notre groupe de référence, à l'exception de nos deux caractéristiques discrètes. Cet individu est donc une femme noire, à éducation inférieure au BAC, ayant un faible revenu, ayant voté pour Kerry aux dernières élections, se considérant libérale, s'identifiant au parti démocrate, qui veut retirer l'armée de l'Iraq au plus vite, qui partage les valeurs de Obama, qui pense qu'il est un bon leader et qui a un âge élevé. Nous voyons donc que cet individu fait partie des groupes les plus discriminés de la société, on retrouve ainsi l'image d'un Obama proche des minorités, qui n'ont souvent que peu de connaissances politiques et votent donc majoritairement par rapport à leurs opinions ou ressentis.

En plus de cette observation globale, nous pouvons aussi nous intéresser à une variable précise, dont les estimateurs fournissent des résultats très pertinents, comme *"Last Vote"*. Les estimateurs de cette variable nous partagent une information importante pour la compréhension de l'élection d'Obama : sa capacité à fédérer des opinions et des individus divers. En effet, la différence de probabilité estimée entre un individu ayant voté pour Kerry en 2004 et un ayant voté pour Bush est de 22 p.p tandis que la différence entre un électeur de Kerry et un n'ayant pas voté est de 10 p.p, ainsi si les abstentionnistes de 2004 votent en 2008 de façon aléatoire ou tout du moins équiprobable, la différence de probabilité estimé devrait être de 11 p.p et non de 10. Cette différence en faveur d'Obama, nous prouve qu'il a su aller chercher des électeurs moins actifs et donc sûrement moins intéressés par la politique, chose qu'il n'a pu faire que grâce aux émotions et opinions.

# III. Analyse du modèle par bloc

Nous nous intéressons ici à l'importance relative de chacun des blocs précédemment définis au travers d'une régression par bloc. En effet, même si quasiment chacune de nos variables est à la fois statistiquement et économiquement significative, il serait intéressant de les comparer en fonction de leur bloc d'appartenance afin de définir leur importance relative. Afin de mener cette étude, nous utiliserons majoritairement deux moyens statistiques, l'analyse de variance et le critère d'information d'Akaike, l'AIC.

Ces deux moyens seront appliqués de façon combinatoire, i.e. pour chaque possibilité de combinaison de bloc, qui sont en les notant respectivement 1, 2 & 3 : {1, 2, 3, 1&2, 1&3, 2&3, 1&2&3}. Ces tests seront de plus effectués sur un même panel cylindré, c'est-à-dire l'intersection des panels sans valeurs manquantes pour chaque combinaison de bloc, ce qui est équivalent au panel sans valeurs manquantes associé à la régression complète.

## A. Analyse de la Variance

L'analyse de la variance explique la différence de la somme des carrés expliqués d'un modèle à un autre. Cette somme représentant la qualité d'ajustement de nos modèles par rapport aux données réelles, plus celle-ci est élevée et donc proche de la somme des carrés totaux, plus le modèle s'ajuste aux données et est donc efficace. Ces sommes de carrés sont plus souvent usitées sous la forme du $R^2$, le coefficient de *"goodness-of-fit"*, qui est calculé ainsi, $\frac{Somme\ des\ carrés\ expliqués}{Somme\ des\ carrés\ totaux}$, ce coefficient appartenant à [0,1] nous indique la part de la variance expliquée par le modèle.

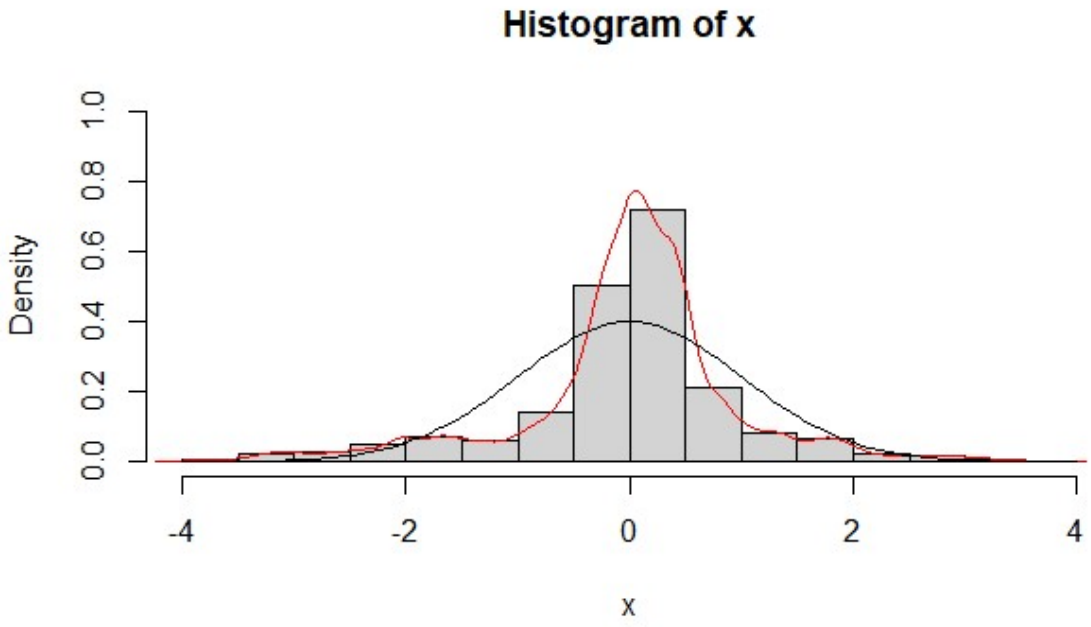
Concernant les degrés de liberté, nous avons un nombre conséquent d'observations dans le panel cylindré, 1591, et l'amplitude des degrés de liberté associés à chaque modèle est de 16. Ainsi, l'impact des degrés de liberté qui participent au calcul du $R^2$ ajusté est ici négligeable ($\frac{16}{1591} = 0.01$).

Concernant les hypothèses nécessaires à la réalisation d'une analyse de la variance, nous nous retrouvons ici dans l'impossibilité de tester celle d'homoscédasticité. En effet, étant donné que nous disposons de 15 variables à plusieurs modalités chacune, la réalisation de groupes correspondant à toutes les combinaisons possibles de modalités nous amène presque sûrement à des groupes vides ou à un individu. Nous avons tout de même essayé d'effectuer sous R des tests de Bartlett et de Levene, mais ceux-ci ne nous ont retourné que des erreurs. Le seul test que nous avons pu effectuer est celui de Shapiro-Wilk, un test de normalité que nous avons appliqué aux résidus de notre régression, et qui nous a naturellement retourné que les résidus ne respectent pas l'hypothèse de normalité[5].

Afin de pouvoir tout de même effectuer l'ANOVA, nous avons voulu tout de même vérifier si nous pouvions dire que nos résidus s'apparentaient au moins un peu à une loi normale. En effet, les tests de normalité ne sont pas robustes et ne tolèrent donc pas des écarts dans les distributions, tandis que l'ANOVA appliquée à un grand échantillon est quant à elle robuste.

[5] Dans le résultat de la commande R *"shapiro.test()"* nous avions comme p-valeur associée à l'hypothèse nulle : "l'échantillon suit une loi normale", une approximation numérique <2.2e-16 i.e. à un seuil infiniment petit, nous pouvons rejeter H0.

Pour ce faire, nous avons centré réduit les résidus du modèle complet puis avons tracé leur histogramme, leur densité (en rouge) et la densité d'une loi normale centrée réduite (en noir).




Nous voyons bien ici que la densité des résidus peut être approximée par une loi normale mais avec un kurtosis beaucoup plus élevé. Nous nous contenterons donc de cette approximation pour effectuer l'analyse de la variance, en se fiant à ses propriétés asymptotiques.

Voici le résultat de la commande *"Anova( )"* de R qui nous permet d'effectuer automatiquement un test comparatif d'analyse de la variance pour nos différents modèles.
On note ici les blocs de cette façon : 1 = *Socio-démographie*, 2 = *Idéologie*, 3 = *Opinion*

```
  Res.Df    RSS  Df Sum of Sq
1   1595 112.85
2   1603 114.01  -8    -1.153
3   1603 136.50   0   -22.496
4   1600 158.11   3   -21.611
5   1611 138.38 -11    19.737
6   1608 162.51   3   -24.135
7   1608 363.69   0  -201.178
---
Analysis of Variance Table

Model 1: 1&2&3  Model 2: 2&3  Model 3: 1&3  Model 4: 1&2  Model 5: 3
Model 6: 2  Model 7: 1
```

Dans cette table nous allons donc nous intéresser à la valeur de la colonne *"Sum of Sq."* qui correspond à la différence de la somme des carrés résiduels de la ligne (n-1) moins celle de la ligne n, cette valeur est liée inversement à la somme des carrés expliqués par la formule :
Somme des carrés expliqués = Somme des carrés totaux- Somme des carrés résiduels.
Ainsi, cette valeur s'interprète aussi comme le delta des carrés expliqués entre les deux modèles.
Plusieurs résultats sont ici intéressants :

- Ligne 2 : ici, le retrait du bloc socio-démographique par rapport au modèle complet n'enlève que 1.15 à la somme des carrés expliqués, ce qui est très faible. On peut donc en conclure que le bloc *Socio-démographie* n'explique pas une grande part de la variance relativement au modèle complet.
- Ligne 5 : ici, le remplacement des blocs socio-démographique et idéologique par celui des opinions nous donne un résultat surprenant. En effet, ici, la différence est positive. Cela signifie que le bloc *Opinion* explique plus la variance de vote que les blocs *Socio-démographie* & *Idéologie* réunis.
- Ligne 6 : ici, relativement au bloc *Socio-démographie*, le bloc *Idéologie* fait perdre 24.135 à la somme des carrés expliqués, ce qui montre une moins bonne capacité d'explication du bloc *Idéologie* par rapport au bloc *Opinion*.
- Ligne 7 : ici, on remarque que par rapport au bloc *Idéologie,* le bloc *Socio-démographie* amène une perte de 201.178 dans la somme des carrés expliqués et donc de - (201.178 + 24.135) = -225.313 par rapport au bloc *Opinion*, ce qui indique un très faible niveau de variance expliquée du premier bloc par rapport aux deux blocs précédents.

De ces résultats quantitatifs, nous pouvons tirer une conclusion qualitative. Tout d'abord, le bloc *Opinion* est le bloc qui explique le mieux le modèle, au point où ce bloc suffit à lui tout seul à fournir une meilleure régression que les blocs *Socio-démographie* et *Idéologie* conjoints. Nous retrouvons ici le résultat attendu dans l'introduction, l'élection d'Obama a été le plus fortement déterminée par les opinions. De plus, nous pouvons aussi remarquer que relativement au bloc *Opinion*, le bloc *Idéologie* explique une part certes plus faible de la variance mais qui reste cependant conséquente. Ainsi, même si Obama s'est fortement appuyé sur les opinions pour son élection, nous ne pouvons tout de même pas dire qu'il a réussi à s'affranchir totalement des considérations idéologiques. Enfin, nous pouvons remarquer que les caractéristiques socio-démographiques ont eu un impact très faible voir quasiment nul sur l'élection d'Obama. Cela nous montre que son électorat est disparate, et qu'il a su toucher par ses prises de positions toutes les classes sociales, toutes les ethnies et tous les âges.

## B. Critère d'information d'Akaike (AIC)

Le critère AIC est un critère de qualité d'un modèle ne reposant pas seulement sur la qualité d'ajustement mais aussi sur le critère de parcimonie[6]. En effet, selon la formule des carrés expliqués, il est mécaniquement possible d'augmenter le $R^2$, autant qu'on le souhaite, en ajoutant des variables aux modèles, peu importe leur pouvoir explicatif. Le AIC permet donc de pénaliser le nombre de variables d'un modèle et se calcule selon la formule suivante :

$$AIC = 2k - 2\ln(L)$$

*Avec k le nombre de variable, et L le maximum de la fonction de vraisemblance du modèle.*

Ainsi, un modèle sera meilleur selon le critère AIC, plus son AIC sera faible.

[6] Le critère de parcimonie, ou *"rasoir d'Ockham"*, est un principe philosophique fondamental en sciences qui prescrit de préférer les modèles d'hypothèses les plus simples lorsque l'on cherche à inférer sur des caractéristiques réelles.

| | K | AICc | Delta_AICc |
|---|---|---|---|
| 1&2&3 | 23 | 330.72 | 0.00 |
| 2&3 | 15 | 330.76 | 0.04 |
| 1&3 | 15 | 621.96 | 291.24 |
| 3 | 7 | 627.78 | 297.05 |
| 1&2 | 18 | 865.74 | 535.02 |
| 2 | 10 | 893.81 | 563.09 |
| 1 | 10 | 2196.38 | 1865.66 |

Voici le résultat de la commande *"AIC( )"* de R qui nous retourne un tableau où les modèles sont triés par AIC croissant, i.e. les meilleurs modèles sont en haut du tableau, et les moins bons, en bas. La colonne K, quant à elle, représente le nombre de variables, et donc la pénalisation de chaque modèle.

Ici aussi, plusieurs résultats sont pertinents pour notre analyse :

- La différence d'AIC entre les deux premières lignes nous montre que malgré l'ajout de 8 variables sur 15 déjà présentes, le bloc 1 apporte tout de même une baisse de l'AIC, ce qui peut donc s'expliquer par une hausse de la vraisemblance.
- La différence d'AIC entre les lignes deux et trois nous montre que relativement au bloc 3, sa combinaison avec le bloc 2 est fortement préférable à celle avec le bloc 1, et ceci pour un nombre identique de variables (i.e. la baisse d'AIC est amenée seulement par une hausse de la vraisemblance).
- La différence d'AIC entre les lignes quatre et cinq nous apporte une information très significative. En effet, selon le critère AIC, il serait préférable de se contenter d'un modèle ne comportant que le bloc 3 plutôt qu'un modèle composé des blocs 1&2.

## Conclusion de l'analyse du modèle par bloc

Ainsi, de ces observations quantitatives, nous pouvons tirer des conclusions qualitatives sur les modèles à préférer, et surtout sur la qualité amenée par chaque bloc. Nous retrouvons évidemment les conclusions de l'analyse de variance, notamment sur l'ordre d'importance des blocs qui sont dans l'ordre décroissant : *Opinion, Idéologie*, *Socio-démographie*. Cependant, ici, nous pouvons apporter plus de nuances et de précision à ces conclusions. Malgré son nombre de variables élevé et sa faible quantité de carrés expliqués, le bloc *Socio-démographie* reste pertinent dans notre analyse. Il n'a certes pas été déterminant dans l'élection d'Obama mais il permet tout de même d'augmenter la vraisemblance de nos modèles statistiques, sûrement grâce à un effet de contrôle. De plus, le bloc *Idéologie* apparaît ici comme un bloc important dans l'analyse de l'élection, même s'il ne l'est pas autant que celui des opinions. Enfin, ce dernier bloc, *Opinion*, renforce encore sa position de dominance en termes de qualité et d'impact.

De ces analyses statistiques, il ressort une version plus nuancée de notre supposition initiale selon laquelle Obama aurait pu s'affranchir des déterminants électoraux classiques au profit des opinions. En effet, même si nous avons pu observer à quel point les opinions avaient

été importantes pour son accession à la présidence, elle a tout de même été fortement influencée par les principes idéologiques des citoyens des Etats-Unis. Et même si les caractéristiques socio-démographiques n'ont certainement pas joué un rôle très important, elles restent tout de même significatives, et importantes dans les études statistiques en leur qualité de variable de contrôle.

# Conclusion

A l'aide des outils économétriques et statistiques, nous avons pu réaliser une analyse ciblée des déterminants de l'élection américaine de 2008, où fût élu le premier président de couleur des Etats-Unis d'Amérique, Barack Obama.

Nous avons, tout d'abord, pu analyser différentes variables recueillies sous la forme d'un questionnaire par la NAES. Les méthodes de la statistique descriptive et de la régression linéaire simple nous ont alors permis de trier et recoder ces variables, afin d'obtenir des modalités en accord avec notre objectif : tester la véracité de l'hypothèse selon laquelle Barack Obama a été le "Président de l'espoir".

Les variables, préalablement traitées, nous ont servi lors d'une régression linéaire multiple, où nous avons déjà pu remarquer l'importance des variables appartenant au bloc *Opinion*. De plus, à l'aide de tests et d'une étude plus théorique, nous avons pu déterminer que notre modèle complet était suffisamment qualitatif pour nous permettre d'en tirer des conclusions économiques.

De cette vision en détail de notre régression et de notre modèle, nous avons décidé d'en adopter une plus globale, en nous intéressant à la théorie des modèles par blocs et des critères de qualité de ceux-ci. C'est ainsi, à l'aide de l'analyse de la variance et du critère d'information d'Akaike, que nous avons pu comparer les différents modèles déterminés par la combinaison de nos blocs. De cette analyse statistique nous sommes arrivés à une conclusion plus nuancée que la précédente, où, même si le bloc *Opinion* s'impose comme le bloc dominant, les deux autres sont tout de même porteurs d'informations et amènent de la qualité à notre travail.

Cependant, malgré des résultats obtenus en adéquation avec la littérature et nos intuitions, nous sommes conscients des limites de notre modèle. Tout d'abord, concernant les données utilisées, notre base de données comporte énormément de valeurs manquantes, comme nous l'avons vu en détail avec la variable du salaire, ce qui réduit la taille de notre panel cylindré.

De même, étant donné le grand nombre de variables et leur caractère subjectif, des hypothèses importantes ne sont pas validées, comme celles de normalité et d'homoscédasticité. Ainsi, une forte part de notre travail se base non pas sur le respect de ces hypothèses, mais sur la robustesse des méthodes employées, et nous exposent à leur principal défaut : la détermination d'un seuil à partir duquel les propriétés asymptotiques sont presque sûrement vérifiées.

Nous sommes aussi conscients que notre modèle à probabilité linéaire n'est sûrement pas convenable en termes de forme fonctionnelle. C'est d'ailleurs pour cette raison que nous avons en premier lieu pensé à un modèle Logit. Cependant, après une étude plus approfondie de celui-ci, nous avons convenu que le remplacement de notre modèle par un Logit n'apporterait pas réellement de valeur ajoutée, et ne rendrait la lecture que plus dure.
En effet, ce type de modèle n'est pas robuste et suppose, dès sa mise en place, que les résidus suivent une loi normale, ce qui n'est absolument pas le cas dans notre modèle.
De plus, pour reprendre les propos de Monsieur Senne, à la lumière de la littérature récente, les modèles Logit ne seraient en général pas meilleurs que ceux à probabilité linéaire et n'apporteraient donc qu'une façon de borner les probabilités estimées entre 0 et 1.

# BIBLIOGRAPHIE

Cours de L3 d'économétrie de la faculté Jean Monnet, J-N. Senne

Cours de 1A de statistiques de l'ENSAE, M. Lerasle & D. Obst

# WEBOGRAPHIE

Article de référence sur lequel nous nous sommes appuyés pour ce papier :
*"Voter Affect and the 2008 U.S. Presidential Election: Hope and Race Mattered", C. Finn & J. Glaser*

- https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.1043.7918&rep=rep1&type=pdf

Partie informatique/statistiques :

- http://perso.ens-lyon.fr/lise.vaudor/non-respect-des-hypotheses-du-modele-lineaire-anova-regression-cest-grave-docteur/
- https://fr.wikipedia.org/wiki/R%C3%A9gression_lin%C3%A9aire
- https://fr.wikipedia.org/wiki/Test_de_Shapiro-Wilk
- https://fr.wikipedia.org/wiki/Crit%C3%A8re_d%27information_d%27Akaike
- https://fr.wikipedia.org/wiki/Analyse_de_la_variance

Partie économie/politique :

- https://perspective.usherbrooke.ca/bilan/servlet/BMEve/1025
- https://fr.wikipedia.org/wiki/Barack_Obama
- https://fr.wikipedia.org/wiki/%C3%89lection_pr%C3%A9sidentielle_am%C3%A9ricaine_de_2008
- https://fr.statista.com/statistiques/559016/revenu-du-menage-aux-etats-unis-repartition-en-pourcentage-en/

# ANNEXE

## 1. Analyse de la distribution des valeurs manquantes de RCa10.

```
Residuals:
     Min       1Q   Median       3Q      Max
-1.08423 -0.15350 -0.08631 -0.02225  1.04237

Coefficients:
                                                                 Estimate Std. Error t value Pr(>|t|)
(Intercept)                                                     3.007e-01  1.163e-01   2.586 0.009778 **
WA01_cmale                                                     -2.256e-02  1.445e-02  -1.561 0.118752
WA02_c                                                          1.515e-04  9.010e-05   1.682 0.092800 .
WA03_cdon't know                                               -1.998e-01  2.347e-01  -0.851 0.394685
WA03_cfour-year college degree                                 -2.003e-02  2.692e-02  -0.744 0.456923
WA03_cgrade 8 or lower                                         -8.849e-02  7.367e-02  -1.201 0.229829
WA03_cgraduate or professional degree                          -6.521e-02  2.802e-02  -2.327 0.020067 *
WA03_cgraduate or professional school after college, no degree -5.530e-02  3.931e-02  -1.407 0.159629
WA03_chigh school diploma or equivalent                         1.660e-03  2.783e-02   0.060 0.952459
WA03_cno answer                                                -3.815e-01  3.353e-01  -1.138 0.255330
WA03_csome college, no degree                                  -3.615e-02  2.814e-02  -1.285 0.198966
WA03_csome high school, no diploma                             -7.511e-03  4.875e-02  -0.154 0.877576
WA03_ctechnical or vocational school after high school         -3.630e-02  5.366e-02  -0.677 0.498769
WA04_c$100,000 to less than $150,000                           -1.167e-01  4.430e-02  -2.634 0.008492 **
WA04_c$15,000 to less than $25,000                             -9.794e-02  4.591e-02  -2.133 0.032997 *
WA04_c$150,000 or more                                         -7.752e-02  4.523e-02  -1.714 0.086727 .
WA04_c$25,000 to less than $35,000                             -7.568e-02  4.642e-02  -1.630 0.103160
WA04_c$35,000 to less than $50,000                             -1.003e-01  4.329e-02  -2.316 0.020665 *
WA04_c$50,000 to less than $75,000                             -9.744e-02  4.236e-02  -2.301 0.021510 *
WA04_c$75,000 to less than $100,000                            -8.269e-02  4.312e-02  -1.917 0.055308 .
WA04_cdon't know                                                8.014e-03  6.127e-02   0.131 0.895942
WA04_cless than $10,000                                        -4.781e-02  5.961e-02  -0.802 0.422594
WA04_cno answer                                                -5.788e-02  5.022e-02  -1.153 0.249218
WC03_casian                                                     1.424e-01  9.449e-02   1.507 0.131998
WC03_cblack, african american, or black hispanic               -4.478e-03  6.860e-02  -0.065 0.947955
WC03_cdon't know                                               -1.340e-01  1.809e-01  -0.740 0.459196
WC03_chispanic, no race given                                   1.356e-02  1.084e-01   0.125 0.900489
WC03_cmixed race                                                8.135e-02  9.665e-02   0.842 0.400042
WC03_cno answer                                                 1.777e-01  9.169e-02   1.938 0.052745 .
WC03_cother                                                     1.762e-01  1.110e-01   1.587 0.112565
WC03_cwhite or white hispanic                                   5.489e-02  6.326e-02   0.868 0.385675
RD01_cdid not vote for president                                9.079e-02  2.894e-02   3.137 0.001727 **
RD01_cdon't know                                                4.380e-02  6.739e-02   0.650 0.515750
RD01_ckerry                                                    -2.963e-02  2.238e-02  -1.324 0.185669
RD01_cnader                                                     2.465e-01  6.379e-02   3.864 0.000115 ***
RD01_cno answer                                                 5.148e-01  9.416e-02   5.467 5.10e-08 ***
RD01_cother                                                     2.573e-01  5.567e-02   4.622 4.02e-06 ***
MA04_cmoderate                                                 -5.949e-02  5.021e-02  -1.185 0.236225
MA04_cno answer                                                 2.857e-01  1.458e-01   1.959 0.050187 .
MA04_csomewhat conservative                                    -5.746e-02  5.165e-02  -1.112 0.266099
MA04_csomewhat liberal                                         -9.809e-02  5.162e-02  -1.900 0.057512 .
MA04_cvery conservative                                        -9.453e-02  5.306e-02  -1.781 0.074976 .
MA04_cvery liberal                                             -1.148e-01  5.510e-02  -2.083 0.037398 *
MA01_cdon't know                                                3.264e-02  5.666e-02   0.576 0.564603
MA01_cindependent                                               7.186e-02  1.960e-02   3.667 0.000252 ***
MA01_cno answer                                                 1.045e-01  8.089e-02   1.292 0.196548
MA01_cother                                                     6.850e-02  4.824e-02   1.420 0.155738
MA01_crepublican                                               -2.785e-02  2.497e-02  -1.116 0.264745
CDb01_ckeep troops until stable government                     -9.858e-02  8.241e-02  -1.196 0.231742
CDb01_cno answer                                               -1.106e-01  1.502e-01  -0.736 0.461613
CDb01_cnone of these                                           -4.188e-02  9.508e-02  -0.440 0.659622
CDb01_cset withdrawal deadline                                 -4.164e-02  8.202e-02  -0.508 0.611731
CDb01_cwithdraw as soon as possible                            -3.353e-02  8.209e-02  -0.408 0.683008
ABo12_c                                                         2.131e-04  5.915e-05   3.602 0.000323 ***
ABo05_c                                                         1.995e-04  6.695e-05   2.979 0.002921 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3261 on 2164 degrees of freedom
  (1518 observations deleted due to missingness)
Multiple R-squared:  0.135,     Adjusted R-squared:  0.1134
F-statistic: 6.253 on 54 and 2164 DF,  p-value: < 2.2e-16
```

Régression du modèle complet sur RCa10_na

**2. Analyse de la multi colinéarité à l'aide de la commande "*vif(reg)*".**

| | GVIF | Df | GVIF^(1/(2*Df)) |
|---|---|---|---|
| WA01_c | 1.071592 | 1 | 1.035177 |
| WA02_c | 1.162075 | 1 | 1.077996 |
| WA03_c | 1.366155 | 2 | 1.081123 |
| WA04_c | 1.406305 | 2 | 1.088980 |
| WC03_c | 1.171847 | 2 | 1.040442 |
| RD01_c | 3.117961 | 3 | 1.208681 |
| MA04_c | 2.168828 | 2 | 1.213546 |
| MA01_c | 2.556992 | 3 | 1.169378 |
| CDb01_c | 1.868468 | 2 | 1.169153 |
| ABo12_c | 4.049547 | 2 | 1.418573 |
| ABo05_c | 2.993381 | 1 | 1.730139 |

Ici les résultats s'interprètent comme ceci : l'absence de multi colinéarité parfaite dans nos variables est vérifié si la valeur de la colonne 4 : GVIF1/(2*Df) est proche de 1 ce qui est le cas ici. (Dans la littérature associée il n'existe pas de consensus sur une valeur seuil à partir de laquelle cette hypothèse n'est plus vérifié, cependant tous les seuils proposés restent supérieurs à 2 ce qui rend les résultats obtenus satisfaisants)

## 3. Vérification que le salaire n'amène pas de biais non-négligeable.

| 2.5 % | 97.5 % | Estimate | Std. Error | t value | Pr(>\|t\|) | |
|---|---|---|---|---|---|---|
| 0.7953729817 | 1.050831442 | 0.9231022 | 0.0651198 | 14.175 | < 2e-16 | *** |
| -0.0205849697 | 0.032915204 | 0.0061651 | 0.0136379 | 0.452 | 0.651289 | |
| -0.0004072534 | 0.001512321 | 0.0005525 | 0.0004893 | 1.129 | 0.258994 | |
| -0.1257230183 | 0.027884188 | -0.0489194 | 0.0391565 | -1.249 | 0.211728 | |
| -0.1300675177 | 0.022594866 | -0.0537363 | 0.0389157 | -1.381 | 0.167521 | |
| -0.2147729934 | -0.050656558 | -0.1327148 | 0.0418355 | -3.172 | 0.001541 | ** |
| -0.1333100742 | -0.029797043 | -0.0815536 | 0.0263869 | -3.091 | 0.002031 | ** |
| -0.2640652939 | -0.175284375 | -0.2196748 | 0.0226314 | -9.707 | < 2e-16 | *** |
| -0.1557130190 | -0.042398416 | -0.0990557 | 0.0288854 | -3.429 | 0.000621 | *** |
| -0.1030052176 | 0.085514095 | -0.0087456 | 0.0480561 | -0.182 | 0.855616 | |
| -0.1632947477 | -0.074828367 | -0.1190616 | 0.0225513 | -5.280 | 1.47e-07 | *** |
| -0.1321850521 | -0.060005361 | -0.0960952 | 0.0183996 | -5.223 | 2.00e-07 | *** |
| -0.1202925212 | -0.045760305 | -0.0830264 | 0.0189993 | -4.370 | 1.32e-05 | *** |
| -0.2211445339 | -0.006811155 | -0.1139778 | 0.0546364 | -2.086 | 0.037127 | * |
| -0.1871962640 | -0.092628379 | -0.1399123 | 0.0241066 | -5.804 | 7.80e-09 | *** |
| -0.1530880270 | -0.066306737 | -0.1096974 | 0.0221217 | -4.959 | 7.85e-07 | *** |
| -0.1049498213 | -0.034796296 | -0.0698731 | 0.0178831 | -3.907 | 9.73e-05 | *** |
| -0.4106057690 | -0.296780516 | -0.3536931 | 0.0290156 | -12.190 | < 2e-16 | *** |
| -0.3486191966 | -0.259451449 | -0.3040353 | 0.0227301 | -13.376 | < 2e-16 | *** |
| 0.0175592881 | 0.031476651 | 0.0245180 | 0.0035477 | 6.911 | 6.94e-12 | *** |

Intervalle de confiance des paramètres d'une régression sans le salaire à gauche, à droite tableau de la même régression sans le salaire mais sur l'échantillon dont les individus n'ont pas de valeur manquante pour la variable du salaire

# CODE R

```
rm(list = ls())
setwd("~/LICENCE 3/SEMESTRE 2/PROJET Obama")
data1 <- read.csv("Obama_data.csv", sep = ",", header = TRUE)

### projet data1 code final Ã rendre
library(doBy)
library(naniar)
library(car)
library(AICcmodavg)
###################################################################### Choix des
variables

##########################################variable dÃ©pendante

data1$RCa10 = recodeVar(data1$RCa10 , c("Yes", "No", "Don't know") , c('1', '0', '0')) #
recodage en variable indicatrice 1(femme)
sum(is.na(data1$RCa10)) ## = 842 valeurs manquantes

##########################################variables indÃ©pendantes
#recodage des rÃ©ponses manquantes ou non pertinentes de notre enquÃªte en valeurs
manquantes

data1 %>% replace_with_na(replace = list(ABo05_c = c(998, 999))) -> data1
data1 %>% replace_with_na(replace = list(ABo12_c = c(998, 999))) -> data1
data1 %>% replace_with_na(replace = list(CDb01_c = c("no answer", "don't know", "none of
these"))) -> data1
data1 %>% replace_with_na(replace = list(RD01_c = c("no answer", "don't know"))) -> data1
data1 %>% replace_with_na(replace = list(MA04_c = c("no answer", "don't know"))) ->
data1
data1 %>% replace_with_na(replace = list(MA01_c = c("no answer", "don't know"))) ->
data1
data1 %>% replace_with_na(replace = list(WA02_c = c(998, 999))) -> data1
data1 %>% replace_with_na(replace = list(WA03_c = c("no answer", "don't know"))) ->
data1
data1 %>% replace_with_na(replace = list(WA04_c = c("no answer", "don't know"))) ->
data1
data1 %>% replace_with_na(replace = list(WC03_c = c("no answer", "don't know"))) ->
data1

#variable par variable(recodage, factor, valeurs manquantes, regression lineaire simple)
########################Bloc socio-dÃ©mographique
## Sex
sum(is.na(data1$WA01_c))##=0
data1$WA01_c <- as.factor(data1$WA01_c)
summary(reg_femme <- lm(data1$RCa10 ~ data1$WA01_c))
## Age
sum(is.na(data1$WA02_c))##=24
```

```
summary(reg_age <- lm(data1$RCa10 ~ data1$WA02_c))
## Education
sum(is.na(data1$WA03_c))##=4
data1$WA03_c = recodeVar(data1$WA03_c,c(unique(data1$WA03_c)),c("Some
education","Lot of education","Some education","Lot of education","No education","No
education","Some education","Lot of education","Some education",NA))
data1$WA03_c <- as.factor(data1$WA03_c)




data1$WA03_c <- relevel(data1$WA03_c, "No education")
reg_educ = lm(data1$RCa10 ~ data1$WA03_c)
summary(reg_educ)
## Salary
sum(is.na(data1$WA04_c)) ##=1662
data1$WA04_c = recodeVar(data1$WA04_c,
c(unique(data1$WA04_c)),c("li","hi","mi","mi","hi","li","mi",NA,"hi","li"))
data1$WA04_c = as.factor(data1$WA04_c)
data1$WA04_c = relevel(data1$WA04_c,"li")
summary(lm(RCa10 ~ WA04_c,data1 ))
## Race
sum(is.na(data1$WC03_c)) ##=43
data1$WC03_c = recodeVar(data1$WC03_c,  c(unique(data1$WC03_c)),  c("black", "white",
"other", "other", "other", "other", "other", NA))
data1$WC03_c = as.factor(data1$WC03_c)
data1$WC03_c <- relevel(data1$WC03_c, "black")
summary(reg_race <- lm(RCa10 ~ WC03_c, data1))

################################Bloc IdÃ©logique
## Last vote
sum(is.na(data1$RD01_c))##=74
data1$RD01_c = recodeVar(data1$RD01_c , c("nader") , c("other")) ## on transforme nader
en other car il n'est ni rÃ©publicain ni dÃ©mocrate
data1$RD01_c = as.factor(data1$RD01_c)
data1$RD01_c = relevel(data1$RD01_c, "kerry")
summary(reg_lv<-lm(RCa10 ~ RD01_c , data1))
## Conservative-Liberal
sum(is.na(data1$MA04_c))##=86
data1$MA04_c = recodeVar(data1$MA04_c, c(unique(data1$MA04_c)),
c("moderate","liberal","liberal","conservative","conservative",NA))
data1$MA04_c <- as.factor(data1$MA04_c)
data1$MA04_c <- relevel(data1$MA04_c, "liberal")
summary(reg_cl <- lm(RCa10 ~ MA04_c, data1))
## Id-Party
sum(is.na(data1$MA01_c))##=116
unique(data1$MA01_c)
summary(reg_id <- lm(RCa10 ~ MA01_c, data1))
```

```
##################################Bloc Opinion
## data1 share values
sum(is.na(data1$ABo12_c)) ##=99
data1$ABo12_c = recodeVar(data1$ABo12_c , c('0', '1', '2', '3', '4', '5', '6', '7', '8', '9', '10') ,
c('No', 'No', 'No', 'No', 'No', 'somewhat', 'somewhat', 'yes','yes', 'yes', 'yes'))
data1$ABo12_c <- as.factor(data1$ABo12_c)
data1$ABo12_c =relevel(data1$ABo12_c, "yes")
summary(reg <- lm(RCa10 ~ ABo12_c, data1))
## Withdraw Troops From Iraq
sum(is.na(data1$CDb01_c)) ##=132
data1$CDb01_c = as.factor(data1$CDb01_c)
data1$CDb01_c = relevel(data1$CDb01_c, "withdraw as soon as possible" )
summary(reg_withdraw<-lm(RCa10 ~ CDb01_c, data1))
## data1 leader
sum(is.na(data1$ABo05_c))##=86
#data1$ABo05_c <- as.factor(data1$ABo05_c)
#data1$ABo05_c<- relevel(data1$ABo05_c, "10")
summary(reg <- lm(RCa10 ~ ABo05_c, data1))

#########################################################Regression Multiple
LinÃ©aire
########################################ModÃ¨le
summary(reg123<- lm(RCa10 ~ WA01_c + WA02_c + WA03_c + WA04_c + WC03_c   +
RD01_c + MA04_c + MA01_c   + CDb01_c + ABo12_c + ABo05_c,data1))

#####################Verification HypothÃ¨ses
### VÃ©rifions que notre Ã©chantillon est alÃ©atoire, i.e. repÃ©sentatif et qu'il est pas
biaisÃ©
data1_data <- read.csv("Obama_data.csv", sep = ",", header = TRUE)
data1_data$RCa10_na[is.na(data1_data$RCa10)] = 1 #on construit une variable indicatrice
qui donne 1 si RCa10 est manquante est 0 sinon
data1_data$RCa10_na[!is.na(data1_data$RCa10)] = 0
regNA = lm(RCa10_na ~ WA01_c + WA02_c + WA03_c + WA04_c + WC03_c   + RD01_c
+ MA04_c + MA01_c   + CDb01_c + ABo12_c + ABo05_c,data1_data )
summary(regNA)
### Histogramme de la distribution des residus
x= reg123$residuals
x = (x-mean(x))/sqrt(var(x))
hist(x,freq = FALSE,ylim=c(0,1) )
lines(density(x),col ='red')
curve(dnorm(x, mean=0, sd=1), from = -4, to = 4,add=TRUE)
### Test de normalite des residus
shapiro.test((reg123$residuals))
### MulticolinÃ©aritÃ© des variables
vif(reg123)

######################Possible biais de sÃ©lection Ã  cause de la variable salaire
## On compare le mÃªme modÃ¨le sur deux Ã©chantillons diffÃ©rents
## Regression avec individus dont salaire valeur manquante mais sans la variable Salary
```

```
reg123 = lm(RCa10 ~ WA01_c + WA02_c + WA03_c + WA04_c + WC03_c   + RD01_c +
MA04_c + MA01_c   + CDb01_c + ABo12_c + ABo05_c,data1)
summary(reg123)
confint(reg123) # intervalles de confiance pour chaque paramÃ¨tre
test = subset(  data1,  !is.na(data1$WA04_c))
## Regression sans individus dont salaire valeur manquante mais sans variable Salary
reg123 = lm(RCa10 ~ WA01_c + WA02_c + WA03_c + WC03_c   + RD01_c + MA04_c +
MA01_c   + CDb01_c + ABo12_c + ABo05_c,test)
summary(reg123)

##########################################################Regression Multiple
LinÃ©aire par Blocs
# On commence par enlever toutes les valeurs manquantes de toutes les variables utilisÃ©es
pour que toutes nos rÃ©gressions soient faites sur le mÃ¨me Ã©chantillon et que l'on puisse
les comparer grÃ¢ce Ã  anova
test = subset(
  data1,
  !is.na(data1$WA01_c) &
    !is.na(data1$WA02_c) &
    !is.na(data1$WA03_c) &
    !is.na(data1$WC03_c) & !is.na(data1$RD01_c)
  &
    !is.na(data1$MA04_c) &
    !is.na(data1$MA01_c) &
    !is.na(data1$CDb01_c) &
    #!is.na(data1$CBb01_c) &
    !is.na(data1$ABo12_c)
  & !is.na(data1$WA04_c) & !is.na(data1$RCa10) &!is.na(data1$ABo05_c))

## Bloc 1 (Socio-dÃ©mographiques) : WA01_c + WA02_c + WA03_c + WA04_c +
WC03_c
## Bloc 2 (Fixe) : RD01_c + MA04_c + MA01_c
## Bloc 3 (Fluide) : CDb01_c +ABo5 c  ABo12_c

reg1 = lm(RCa10 ~ WA01_c + WA02_c + WA03_c + WA04_c + WC03_c, test)
reg2 = lm(RCa10 ~ RD01_c + MA04_c + MA01_c, test)
reg3 = lm(RCa10 ~ CDb01_c + ABo12_c + ABo05_c , test)
reg12 = lm(RCa10 ~ WA01_c + WA02_c + WA03_c + WA04_c + WC03_c    + RD01_c +
MA04_c + MA01_c, test)
reg13 = lm(RCa10 ~ WA01_c + WA02_c + WA03_c + WA04_c + WC03_c    + CDb01_c +
ABo12_c + ABo05_c, test)
reg23 = lm(RCa10 ~ RD01_c + MA04_c + MA01_c    + CDb01_c + ABo12_c + ABo05_c,
test)
reg123 = lm(RCa10 ~ WA01_c + WA02_c + WA03_c + WA04_c + WC03_c   + RD01_c +
MA04_c + MA01_c   + CDb01_c + ABo12_c + ABo05_c,test )
summary(reg123) ##Résumé de la régression du modèle complet

anova(reg123, reg23, reg13, reg12, reg3, reg2, reg1) ##Analyse de la variance
aictab(list(reg1, reg2, reg3, reg12, reg13, reg23, reg123),c("1", "2", "3", "1&2", "1&3",
"2&3", "1&2&3")) ##Analyse des modèles selon l'AIC
```