Méthodes quantitatives avancées tp1

Hategekimana

04-04-2022

L'effet du sexe et de la classe d'âge sur le positionnement politique

Base théorique

Dans ce petit rapport, nous allons nous intéresser au lien entre le positionnement politique « Gauche - Droite » et les deux facteurs sociodémographiques que sont le sexe et l'âge. Pour réaliser cette analyse, nous allons utiliser les données du panel suisse des ménages.

L'alignement politique possède plusieurs échelles pour la mesurer, mais la plus connue reste l'échelle « Gauche – Droite ». Cette échelle est censée encapsuler un certain nombre de thématiques (positionnement sur la politique économique, sur la politique sociale, sur la politique migratoire, etc.). L'avantage de cette mesure, est qu'elle est facilement compréhensible en dehors du monde académique. Cependant, étant une dimension unique regroupant un nombre important de thématique, elle n'est pas parfaite et pourrait grandement être améliorée. Mais pour notre rapport, cette échelle est suffisante pour élaborer nos analyses. Ici cette échelle va de « Gauche » (codé 0) à « Droite » (codé 10) avec les nombre allant de 1 à 9 entre les deux. Donc l'échelle (prise dans le sens croissant dans les modèles de régressions) montre une tendance vers le vote à droite. Notons que dans cet exemple « 5 » représente le centre de l'échelle.

Ici le positionnement politique est considéré comme notre variable dépendante, c'est-à-dire, l'élément que nous allons tenter d'expliquer à l'aide d'un modèle statistique et d'autres variables dites explicatives. Ces dernières sont les suivantes : âge et sexe. Ces deux variables sont très importantes, car elles sont des facteurs structurants de la vie d'un individu et cela peu important dans quel milieu social . Voilà pourquoi elles sont utilisées comme variables de contrôle dans l'immense majorité des cas (Marquis, 2006). Ici, nous leur donnons un rôle central, bien que nous n'utilisons pas de variable de contrôle. De plus, ces deux facteurs ont un effet important sur la participation politique des individus, ce sont même de puissants moteurs à la participation politique (Mazzoletti & Masulin, 2005; Sciarini et al., 2001).

Attention toutefois aux simplifications. En ce qui concerne le sexe par exemple, ce n'est pas simplement l'appartenance à un sexe biologique qui détermine l'activité politique, c'est les déterminants sociaux qui sont liés à ce sexe qui ont une grande force d'explication (Alvarez & Parini, 2005). De la même manière pour l'âge. Ce n'est pas l'appartenance à une âge biologique qui déterminent le positionnement politique. Par exemple, ce n'est pas en vieillissant que les individus développent des positions plus à droite. C'est en fait un effet de cohorte qui amène le changement de positionnement politique (Tiberj, 2009). Cependant,

nous ne pourrons pas prendre en compte toute cette complexité en compte dans ce travail, mais il est important de relever ces éléments.

Hypothèses

Dans ce travail, nous avons deux hypothèses principales. Premièrement, le sexe exerce une influence sur le positionnement politique. Dans cette perspective, nous pensons qu'être une femme à une relation négative avec notre échelle de positionnement politique, c'est-à-dire, qu'être une femme diminue les chances d'être à droite. Nous nous attendons donc à avoir un coefficient négatif dans nos modèles de régression.

Deuxièmement, l'âge a un effet positif sur notre échelle. Autrement dit, plus l'âge est avancé, plus les personnes auront tendance à être à droite. Nous nous attendons donc à un coefficient positif. Dans notre cas, nous divisons notre variable d'âge en 5 groupes "16-30", "31-50", "51-70", "71-90" et "91-98". Nous utilisons le groupe "16-30" comme catégorie de référence pour la comparer aux autres. Nous nous attendons à des coefficients positifs pour toutes les autres catégories.

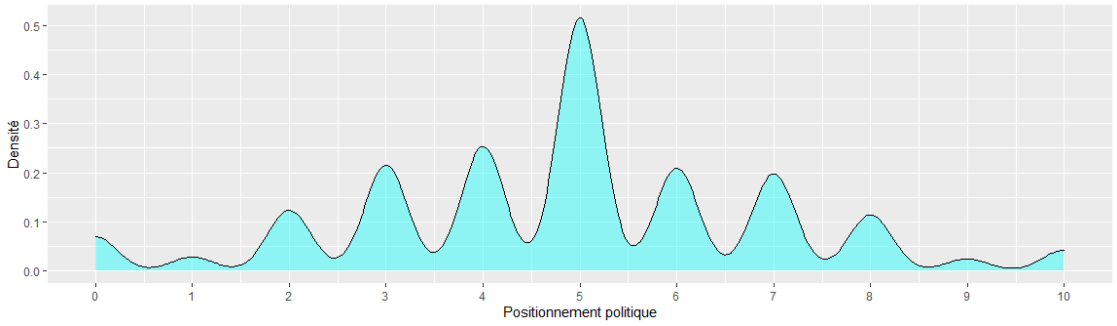
Présentation des modèles

Nous allons utiliser deux modèles pour ce travail: une régression linéaire et une régression ordinale. Nous utiliserons ces deux modèles pour pouvoir les comparer.

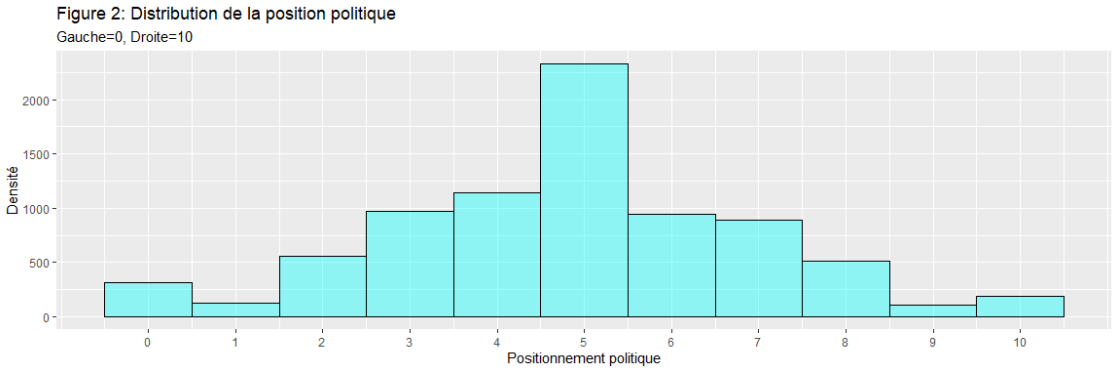
En ce qui concerne la régression linéaire, nous l'utilisons puisqu'à priori nous avons un variable métrique (continue) pour le positionnement idéologique. Classiquement, c'est le modèle qui semblerait le mieux correspondre à la spécificité de notre variable dépendante. Toutefois nous devons souligner deux éléments qui remettent en question ce premier point de vue.

Premièrement, la distribution de la variable, bien qu'elle semble suivre une distribution normale, est discrète: elle ne s'arrête que sur un nombre déterminé d'éléments (de 0 à 11). Voilà pourquoi nous obtenons un résultat étrange lorsque nous essayons de représenter la distribution en utilisant la densité qui est généralement utilisée sur les variables continues.

Figure 1: Distribution de la position politique
Gauche=0, Droite=10



Lorsque nous utilisons un histogramme, nous pouvons voir plus clairement les différentes valeurs que la variable peut avoir. Cette variable est bien discrète puisqu'il n'existe que des entiers pour cette variable.

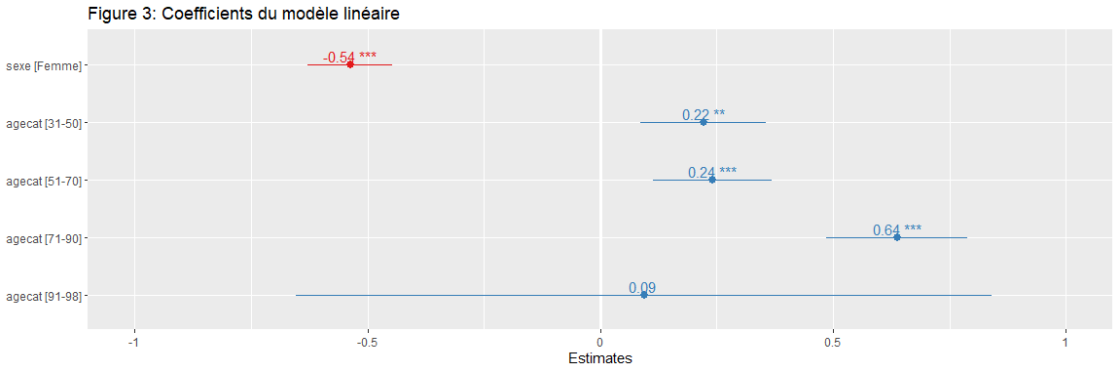


De plus, nous commettons deux erreurs supplémentaires si nous considérons que cette variable est continue. Premièrement, nous utilisons une échelle qui est bornée entre zéro et dix, alors que la régression linéaire n'a pas de limite (va de -∞ à ∞). Deuxièmement, étant une échelle subjective (sans forcément de possibilité pour les individus qui répondent à la question de pouvoir se baser objectivement sur les observations), l'écart entre chaque niveau n'est pas forcément le même. Pourtant la régression linéaire part du principe que les écarts sont les mêmes d'un niveau à un autre de l'échelle. Afin de pouvoir régler ces problèmes, nous devons recourir à une régression ordinale (ou régression logistique ordonnée). Cette régression a l'avantage de prendre en considération les différences d'écart entre les variables. Cependant, les coefficients ne sont pas entièrement interprétables sans transformation.

Nous allons dans un premier temps présenter les résultats de la régression linéaire, puis dans un second temps celles de la régression ordinale.

Régression linéaire

Le graphique suivant permet de voir les coefficients de la régression (points) avec leur intervalle de confiance à 95% (ligne). Un intervalle de confiance à 95% indique l'amplitude de valeurs que le coefficient peut prendre pour lesquelles nous avons une grande confiance. Dans le cas d'une régression linéaire, lorsque l'intervalle de confiance contient la valeur zéro, le coefficient n'est pas significatif. Visuellement parlant, cela veut dire que la ligne du coefficient croise la ligne du zéro de l'axe des abscisses. Comme pour les tableaux de régression classique, la présence d'étoile indique que le coefficient est significatif selon différents seuil (*** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$). Plus il y a d'étoiles, plus nous pouvons être confiant du résultat. Notons encore une dernière chose. Dans ce graphique, il n'y a pas de constante (intercept). Nous avons décidé de ne pas le montrer car ce sont les valeurs des coefficients qui nous intéressent. Voici, ce que donne le graphique:



Nous pouvons voir que tous nos coefficients sont significatifs, excepté la dernière catégorie d'âge ("91-98") qui a un très grand intervalle de confiance. Cela s'explique par la petite taille de cette catégorie puisqu'elle ne contient que 30 personnes (voir annexe).

Nous pouvons remarquer que le coefficent du sexe est significatif. Dans notre cas, nous avons testé l'effet d'être une femme comparé à un homme. Le résultat montre un coefficient négatif, c'est-à-dire qu'être une femme diminue la propension à être à droite de -0.54 sur notre échelle. Ce qui valide notre première hypothèse.

Nous voyons également que pour toutes les catégories d'âge significative, nous avons des résultats positifs. Puisque chaque catégorie est comparée à la catégorie de référence ("16-30"), nous pouvons dire que le fait d'être plus âgé augmente le fait de se trouver à droite dans l'échelle des positions politiques. Étant donné le fait que les coefficients augmentent avec l'ancienneté de la catégorie d'âge, nous pouvons confirmer notre deuxième hypothèse.

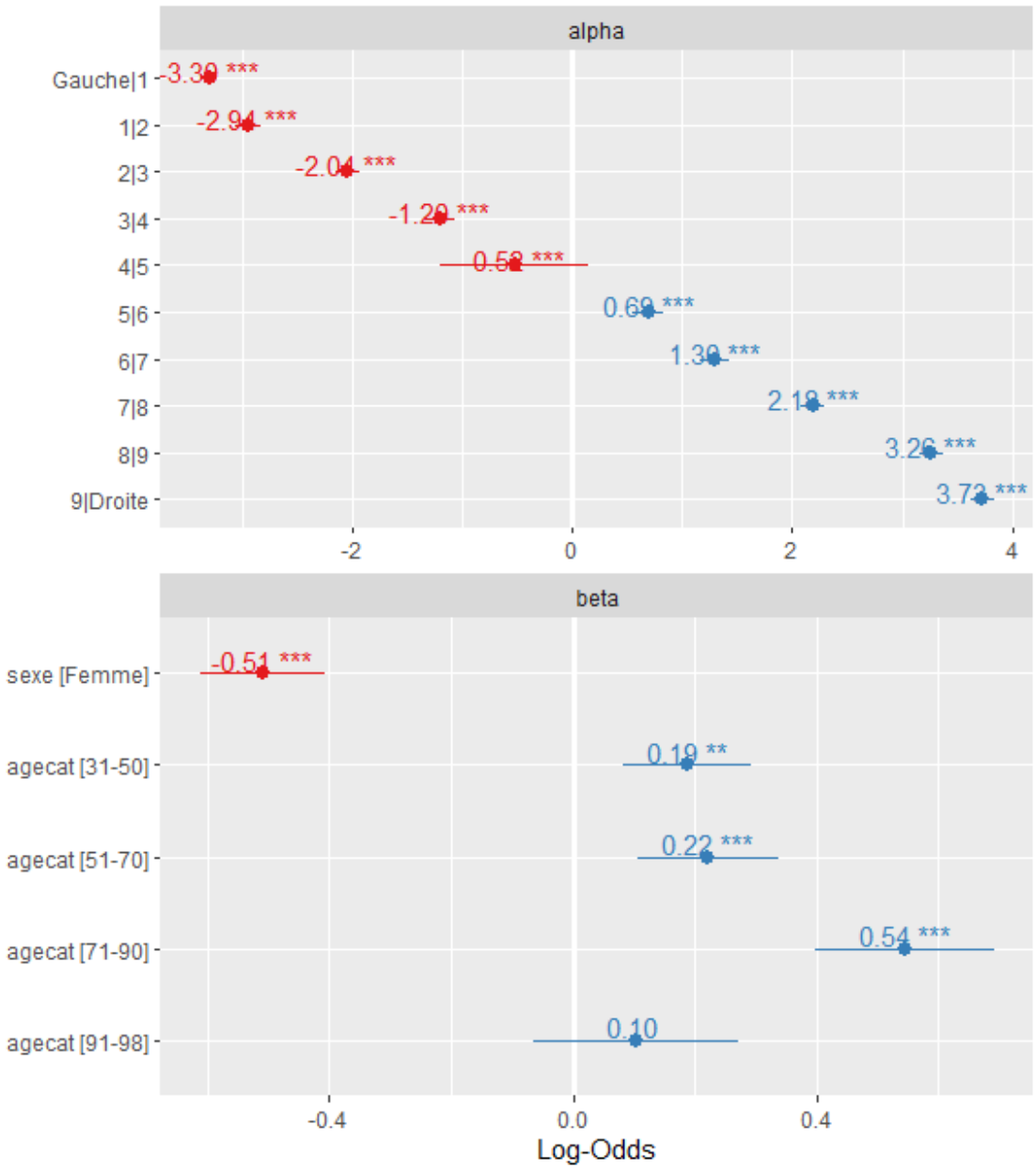
Cependant, pour toutes les raisons que nous avons citées dans la partie sur la description des modèles, nous devons considérer l'utilisation d'un modèle ordinal.

Régréssion ordinale

Le graphique suivant est similaire à celui que nous avons présenté dans la partie sur la régression linéaire. Cependant, il présente quelques particularités. Premièrement, tous les coefficients sont en log-odds qui est le résultat d'une transformation logit (dans les modèle binomiale/logistique, ordinale, multinomiale, etc.). Il faut simplement comprendre que, contrairement à la régression linéaire, nous ne pouvons pas directement interpréter les valeurs puisqu'elles ne sont plus dans la même unité que la variable dépendante. Dans notre situation ce n'est pas un grand problème, puisque l'interprétation d'une échelle n'est déjà pas aisée rendant les signes bien plus importants que les valeurs.

Deuxièmement, il y a une case alpha qui apparaît avant. Elle permet de voir les différences d'écart entre les différents niveaux de notre variable dépendante. Nous pouvons voir que l'écart entre les différentes valeurs n'est pas constant et que l'écart le plus grand se constate entre les valeurs "4" et "5". Ce qui permet de confirmer l'importance d'utiliser un modèle ordinal pour estimer ce modèle.

Figure 4: Coefficients du modèle ordinal



Outre le fait que les coefficients ne sont pas directement interprétables et qu'ils ne sont par conséquent pas comparable avec ceux de la régression linéaire (mais pourtant très proches), nous obtenons les mêmes résultats qu'avec la régression linéaire. Nous avons les mêmes signes, la même hiérarchie entre les coefficients, les mêmes significations et donc les mêmes conclusions.

Nous allons ensuite présenter les coefficients des deux modèles, les uns à côtés des autres pour avoir une meilleure vue de la situation.

Tableau de régression

Dans le tableau de régression qui suit, nous avons nos deux modèles de régression. Nous pouvons constater que pour le modèle linéaire nous avons une constante (intercept), là ou pour le modèle ordinal, nous avons la différence entre les différents niveaux. En dehors des différences citées précédemment, les résultats des coefficients sont assez similaires. Cependant, n'étant pas dans la même unité, ils ne sont pas comparables.

Table 1:Tableau 1: Test des deux modèles de régression

	Linéaire	Ordinale
(Intercept)	4.880 ***	
	(0.058)	
sexeFemme	-0.538 ***	-0.509 ***
	(0.046)	(0.040)
agecat31-50	0.221 **	0.187 **
	(0.069)	(0.059)
agecat51-70	0.242 ***	0.221 ***
	(0.065)	(0.056)
agecat71-90	0.637 ***	0.545 ***
	(0.077)	(0.067)
agecat91-98	0.094	0.103
	(0.381)	(0.341)
Gauche|1		-3.295 ***
		(0.074)
1|2		-2.941 ***
		(0.067)
2|3		-2.039 ***
		(0.057)
3|4		-1.201 ***

		(0.053)
4|5		-0.523 ***
		(0.052)
5|6		0.690 ***
		(0.052)
6|7		1.301 ***
		(0.054)
7|8		2.184 ***
		(0.059)
8|9		3.259 ***
		(0.075)
9|Droite		3.725 ***
		(0.086)
N	8096	8096.000
R2	0.025	
logLik	-17361.928	-16726.761
AIC	34737.857	33483.522

*** p < 0.001; ** p < 0.01; * p < 0.05.

Généralement, le modèle linéaire jouit d'un avantage dans sa lisibilité puisqu'il est simple à interpréter. Effectivement, les coefficients du modèle restent dans l'unité de la variable dépendante ce qui permet une lecture incrémentale aisée. Par exemple nous pouvons facilement dire: "Une femme, lorsque nous contrôlons pour la catégorie d'âge, a en moyenne un score inférieur de -0.538 points dans l'échelle de position politique". Pourtant, ce n'est pas un réel avantage ici pour deux raison:

Premièrement, l'unité dans la variable dépendante n'est pas si pratique à comprendre (qu'est-ce que ça veut dire "-0.538" sur une échelle de position politique? Ce n'est pas clair). Dans notre cas, l'amplitude n'a donc pas vraiment d'importance et le log-odds n'a donc pas de désavantage puisque nous pouvons alors faire le même type d'interprétation.

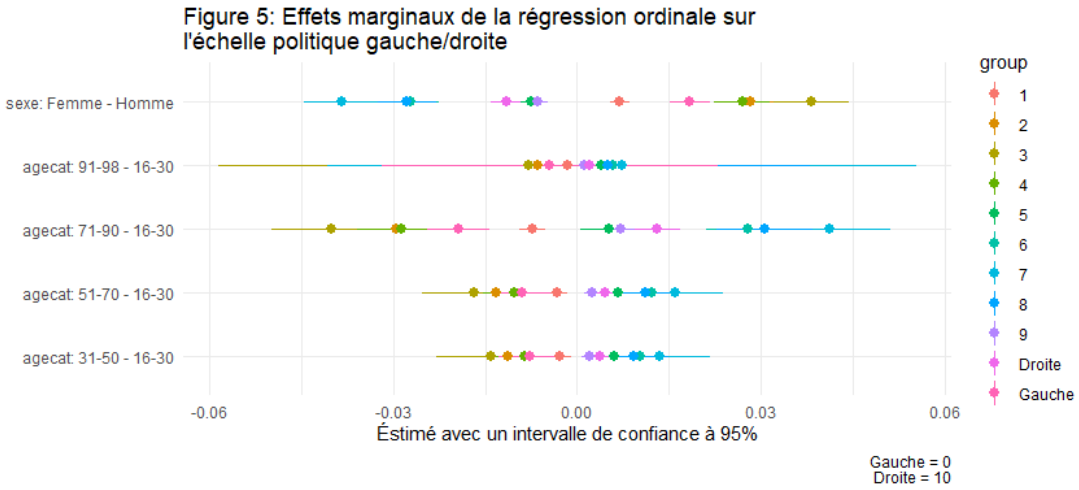
Deuxièmement, les coefficients sont particulièrement petits (petite taille d'effet) dans notre modèle. Effectivement, aucun de ces coefficient n'est supérieur à 0.6, alors que notre variable dépendante ne compte que des entiers. C'est-à-dire que pour avoir un changement qui ait un vrai effet (au moins un), il faudrait combiner les deux variables (par exemple en partant comme catégorie de référence d'une femme entre 16 et 30 ans, il faudrait être un homme entre 71 et 90 ans pour avoir une différence d'au moins un) puisqu'à elles seules elles n'y arrivent pas. Ceci s'explique par le petit R2: le modèle n'explique de 2.5% des différences (de la variance). Ce qui encourage le fait d'utiliser des termes comme "plus de chance"/"moins de chance", ce qui est la spécialité des modèles logit et par extension ordinales.

Si nous ajoutons à ces deux observations, les avantages du modèle ordinal sur le modèle linéaire, nous pouvons alors préférer le premier modèle pour notre analyse. Dans ce qui suit, nous allons regarder en détail ce que donne le résultat du modèle de régression ordinale.

Effets marginaux

L'utilisation des effets marginaux dans un modèle logit permet d'estimer des probabilités prédites pour les coefficients. Un effet marginal est défini comme une mesure de l'association entre un changement dans une variable indépendante et un changement dans la variable dépendante tout en prenant en considération les autres variables indépendantes du modèle (ici on peut aussi remplacer variable indépendante par variable explicative). Dit simplement, quel est l'effet sur la variable dépendante de faire varier une unité de la variable explicative en prenant en compte les autres variables explicatives. Il existe deux méthodes pour le faire, l'average marginal effect (AME) qui est la moyenne de tous les effets marginaux de chaque observation et le marginal effect at the mean (MEM) ou les effets marginaux sont calculés en se basant sur une observation hypothétique qui correspond à la moyenne de toutes les observations. Dans notre cas, nous utiliserons l'average marginal effect (AME) pour estimer nos probabilités prédites.

Une fois que nous avons effectué les calculs, nous obtenons l'image suivante. Pour chaque coefficient, nous obtenons plusieurs points de couleur représentant les diférents niveau de la variable dépendante qui indiquent sur l'axe x, la probabilité prédite avec un intervalle de confiance à 95% pour chaque point (probabilité). Lorsque l'intervalle de confiance croise le zéro de l'axe x, il ne peut pas être considéré comme significatif. Un point à droite du zéro indique un probabilité positive ("plus de chance de"), un point à gauche représente une probabilité négative ("moins de"). Plus un point est éloigné du zéro, plus l'effet de la probabilité est grand.



Par exemple, le fait d'être une femme par rapport à un homme diminue d'environ 4% (-0.04) les chances d'être à "8" sur l'echelle politique. Inversement être une femme par rapport à un homme augmente d'environ 4% (0.04) les chances d'être à "3" sur l'echelle politique. L'avantage de prendre en considération les effets marginaux, est que cela permet de mettre de côté l'hypothèse de linéarité parfaite dans les données. Par exemple, les femmes ont plus de chance de se retrouver à "0" qu'à "1", mais ont plus de chance de se retrouver à "2" qu'à "1".

Ce type de graphique utilise une méthode de comparaison. Dans les exemples que nous avons donnés, les probabilités des femmes sont toujours prises en se référant aux hommes. La visualisation pouvant être assez confuse à comprendre, il existe d'autres méthodes de visualisations. Voici un exemple qui permet de voir les probabilité prédite de chaque catégorie, les unes à côté des autres. Chaque cas représente un des niveaux de la variable dépendante (Gauche=0, x1=1, x2=2, …, x9=9, Droite=10). Il est donc aussi possible de constater ici qu'il n'y a pas de linéarité parfaite. Nous pouvons aussi constater que nous pouvions nous y attendre, la majorité des gens ont indiqué "5" dans l'échelle.

Figure 6: Probabilité prédite du modèle ordinale pour le Sexe

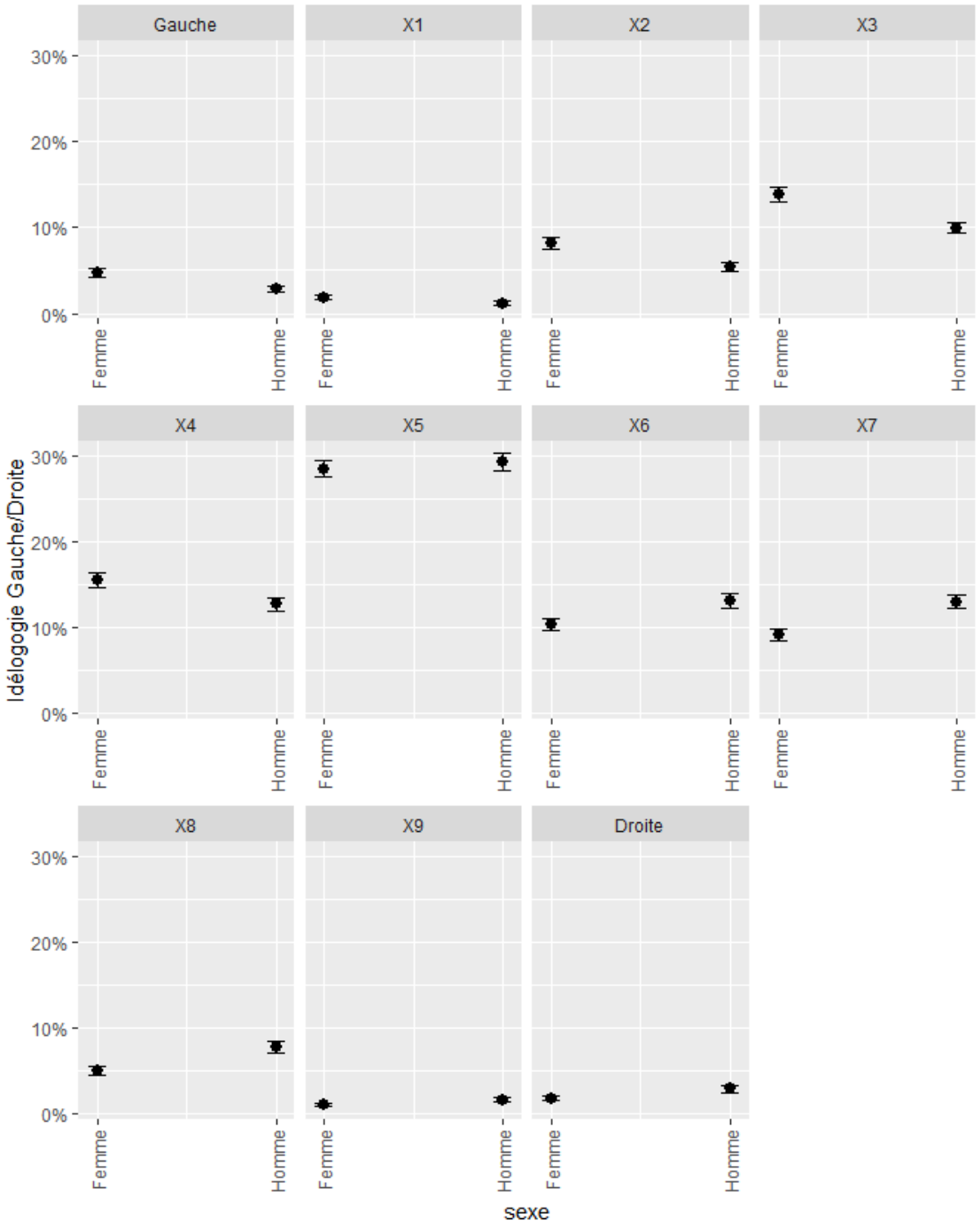
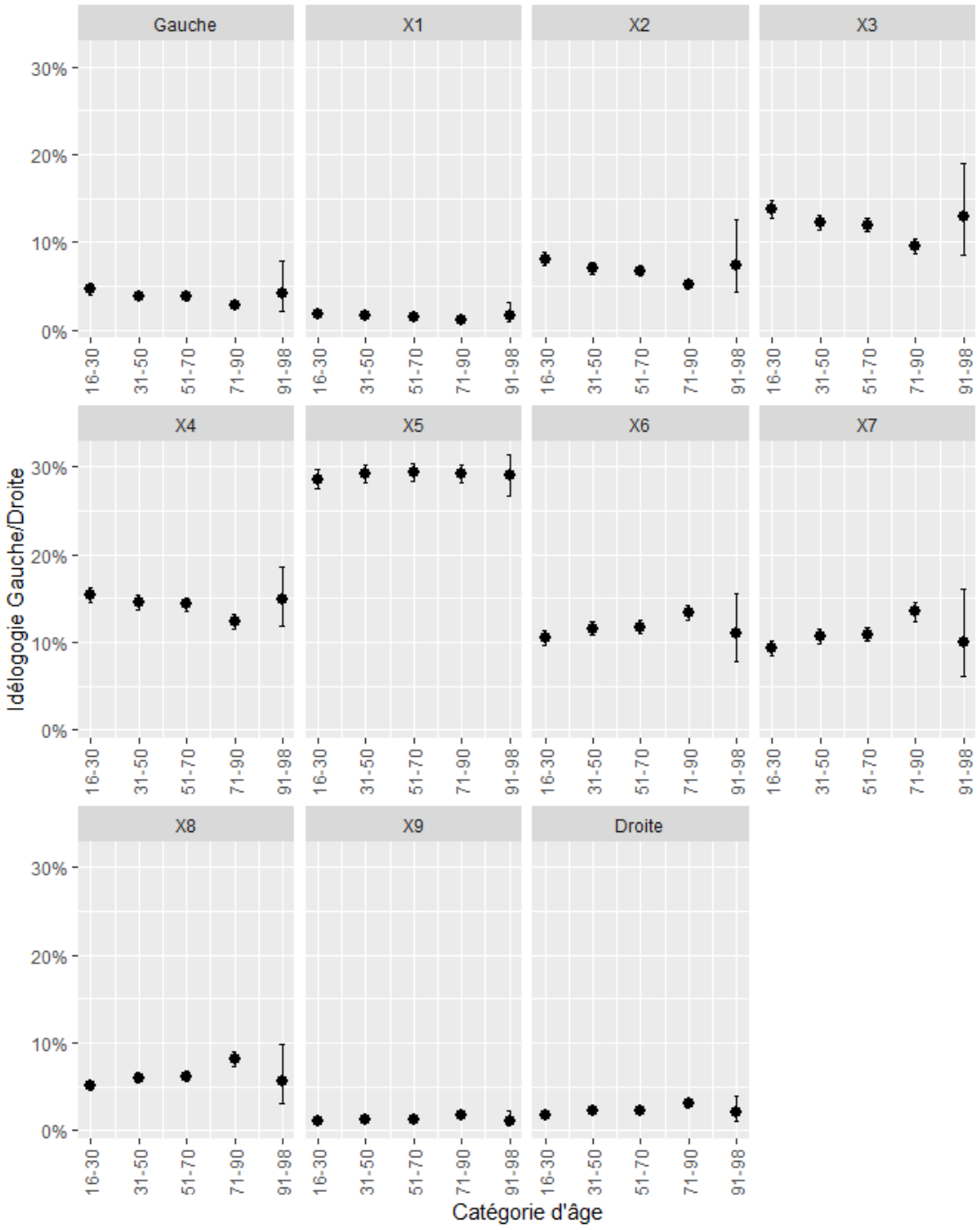


Figure 7: Probabilité prédite du modèle ordinale pour la catégorie d'âge



Conclusion

Nous avons pu confirmer nos hypothèses concernant le lien entre nos variables indépendantes avec notre variable dépendante. Premièrement, être une femme a tendance à diminuer le score de l'echelle de position politique. Deuxièmement, l'ancienneté a tendance à augmenter le score.

Nous avons utilisé deux méthodes pour l'analyse. La régression linéaire classique et la régression ordinale. Bien que la première méthode permet une interprétation facilité des résultats (explication incrémentale de l'effet et conservation de l'unité de mesure de la variable dépendante), elle souffre d'un problème de spécification. Effectivement, ses hypothèses de base font qu'elle ne considère pas la distance entre les différents niveaux de notre variable dépendante, ce qui peut apporter un grand nombre de biais. Dans les faits, les coefficients et les erreurs standards étaient assez similaires. De plus, les mêmes coefficients passaient le seuil de significativité de 0.05 de l'analyse fréquentiste. Mais pour la raison expliquée plus tôt, nous avons plus de confiance dans les résultats du modèle ordinal.

Un point qu'il est intéressant de relever c'est qu'au niveau de l'interprétation, le modèle ordinal n'a pas apporté plus de difficulté par rapport au modèle linéaire. Effectivement, l'unité de mesure de la régression ordinale étant le log-odds qui n'est pas directement interprétable dans l'unité de la variable dépendante, l'interprétation est généralement bien plus sommaire que dans une régression linéaire. Ce ne fut pas le cas dans ce petit rapport. Il y a plusieurs explications à cela. Premièrement, l'unité de mesure de la variable dépendante était assez floue rendant les coefficients de la régression ordinale aussi utile que ceux de la régression linéaire. Deuxièmement, les coefficients des variables indépendantes étaient faibles et ne permettaient pas d'avoir une explication intéressante.

Un autre point qu'il est important de souligner est que l'utilisation des effets marginaux a permis une lecture plus riche et intéressante des résultats de la régression ordinale (dans une régression linéaire, il n'y aurait pas eu de nouvelles informations). Ces effets marginaux ont aussi permis de souligner la non-linéarité des niveaux.

Finalement, ce rapport a permis en plus de confirmer nos résultats, de mettre en avant l'importance de la régression ordinale dans l'analyse statistique.

Annexe

Résumé statistique

Data summary

Name	Piped data
Number of rows	13751
Number of columns	3
_____	
Column type frequency:	
numeric	3
_____	
Group variables	None

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
Position politique	5655	0.59	4.86	2.09	0	4	5	6	10	
Sexe	0	1.00	1.51	0.50	1	1	2	2	2	
Âge	22	1.00	43.86	23.32	0	24	47	63	99	

Données manquantes

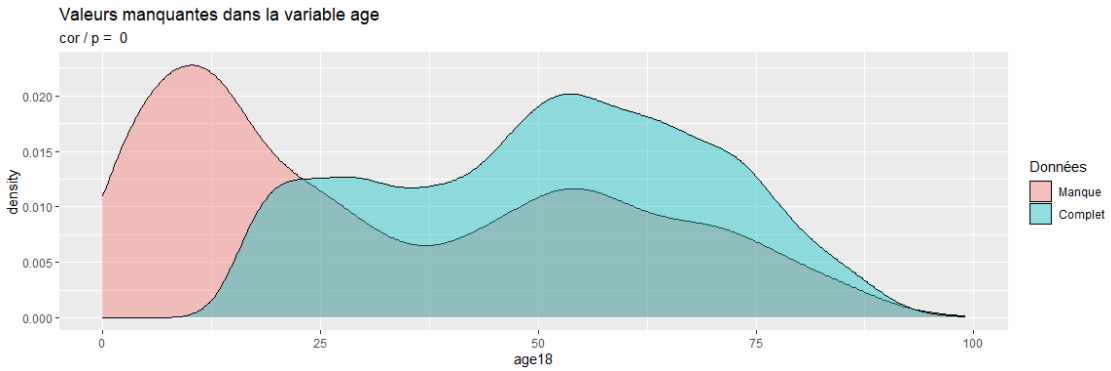
Nous allons brièvement réaliser une analyse des données manquantes. Puisque la variable sur le vote représente l'essentielle des valeurs manquante au point où il n'existe pas de valeur dans cette variable pour lesquelles les deux autres variable (sexe et age) aient des données manquante. Il n'est pas possible de conduire une analyse de données manquantes avec. Voilà pourquoi nous commençons avec le sexe.

Données	Sexe	Freq
Complet[1]	Homme[1]	-1.989218[1]
Manque[1]	Homme[1]	1.989218[1]
Complet[1]	Femme[1]	1.989218[1]
Manque[1]	Femme[1]	-1.989218[1]

[1]Chi2 p = 0.049

Nous voyons dans le tableau des résidus standardisé que le test de Chi2 est significatif indiquant un lien entre le sexe et les variables manquantes (p=0.049). De plus, la distribution des résidus indique une surreprésentation des hommes dans les données manquantes.

Nous pouvons ensuite nous occuper de l'âge.



Le graphique montre que le résultat du test de corrélation entre la variable âge et la variable des données manquantes qu'il existe un lien entre les deux (p=0). De plus, il semblerait que les jeunes sont surreprésentés dans les variables manquantes.

Nous voyons finalement que les jeunes et les hommes sont surreprésentés dans les données manquantes.

Bibliographie

Alvarez, Elvita, and Lorena Parini. 2005. "Engagement Politique Et Genre : La Part Du Sexe." *Nouvelles Questions Féministes* 24 (3): 106–21. https://doi.org/10.3917/nqf.243.0106.

Marquis, Lionel. 2006. *La Formation de l'opinion Publique En Démocratie Directe: Les Référendums Sur La Politique Extérieure Suisse 1981-1995*. Seismo.

Mazzoletti, Oscar, and Maurizio Masulin. 2005. "Jeunes, Participation Politique Et Participation Sociale En Suisse. Une Étude de Cas." *Swiss Political Science Review* 11 (2): 55–81. https://doi.org/10.1002/j.1662-6370.2005.tb00355.x.

Sciarini, Pascal, Than-Huyen Ballmer-Cao, and Romain Lachat. 2001. "Genre, Âge Et Participation Politique: Les Élections Fédérales de 1995 Dans Le Canton de Genève." *Swiss Political Science Review* 7 (3): 81–96.

Tiberj, Vincent. 2009. "L'impact Politique Du Renouvellement Générationnel. Une Comparaison Franco-Américaine." *Agora Débats/Jeunesses* 51 (1): 125–41. https://doi.org/10.3917/agora.051.0125.