UNIVERSITE GASTON BERGER
L'excellence au service du développement
UFR des Lettres et Sciences humaines
Département de sociologie
Semestre 2

SOCIO532.2
STATISTIQUES & INFORMATIQUE APPLIQUÉES AUX SCIENCES SOCIALES
© El Hadj Touré, Ph.D.

LABO SPSS # 7
Régression logistique

La régression linéaire multiple permet d'analyser l'effet de deux variables indépendantes (VIs) quantitatives/dichotomiques ou plus sur une variable dépendante (VD) quantitative. Qu'en est-il lorsque la VD est dichotomique ou dichotomisée? Dans ce genre de problème de recherche, on utilise la régression logistique. La régression logistique comporte des avantages majeurs par rapport à la régression linéaire : 1) elle permet de modéliser la **probabilité** qu'un évènement survienne sous certaines conditions ; 2) elle permet de modéliser les **chances** d'appartenir à un groupe si les individus présentent telles caractéristiques ou autres ; 3) elle ne nécessite pas que la relation soit linéaire ou que la VD suive une distribution normale. Lorsqu'une relation est curvilinéaire ou que la VD quantitative suit une distribution asymétrique, on peut dichotomiser la VD en 0/1 et utiliser ainsi la régression logistique.

De surcroît, lorsqu'une VD est qualitative non dichotomique (j≥3), on la recode de façon à obtenir une variable binaire 0/1. Comme la régression linéaire, la régression logistique peut inclure des VIs dichotomiques ou des variables factices codées 0/1. Nous verrons comment transformer ***automatiquement*** une variable qualitative, nominale ou ordinale, en variables factices, à l'aide de SPSS lorsqu'il s'agit de procéder à l'analyse de régression logistique. Qu'il s'agisse de dichotomiser une VD qualitative ou de créer des variables factices à partir d'une VI qualitative, il faut s'assurer que les catégories sont mutuellement exclusives, que leurs proportions ne sont pas très déséquilibrées, et surtout de bien identifier la *catégorie de référence*.

Dans ce labo SPSS, nous apprendrons à procéder à une analyse de régression logistique simple et multiple, de façon à :

- ✓ Déterminer la signification statistique et réelle de l'effet d'une ou de plusieurs VIs sur une VD dichotomique,
- ✓ Prédire les chances d'appartenir à un groupe et la probabilité qu'un évènement survienne sous certaines conditions,
- ✓ Déterminer la proportion expliquée dans la variation du groupe d'appartenance,
- ✓ Calculer le taux de classement des cas en fonction des valeurs prédites.

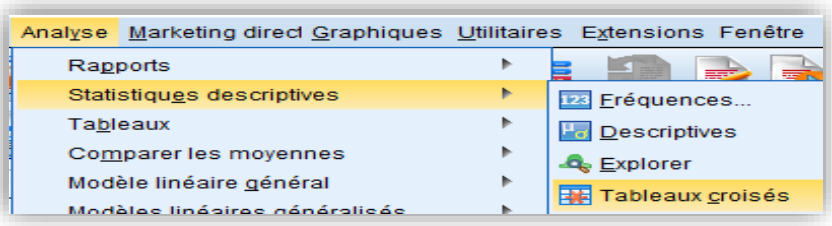
1. Tableau croisé et rapport de cotes

Prenons maintenant un exemple de recherche et considérons la base de données « Sondage_EtudiantsSocioL2_2021(Labo7) ». Intéressons-nous à la relation entre l'**intérêt pour la politique** (interet2) et l'**intention de vote** à la prochaine élection présidentielle (vote) chez les étudiants.

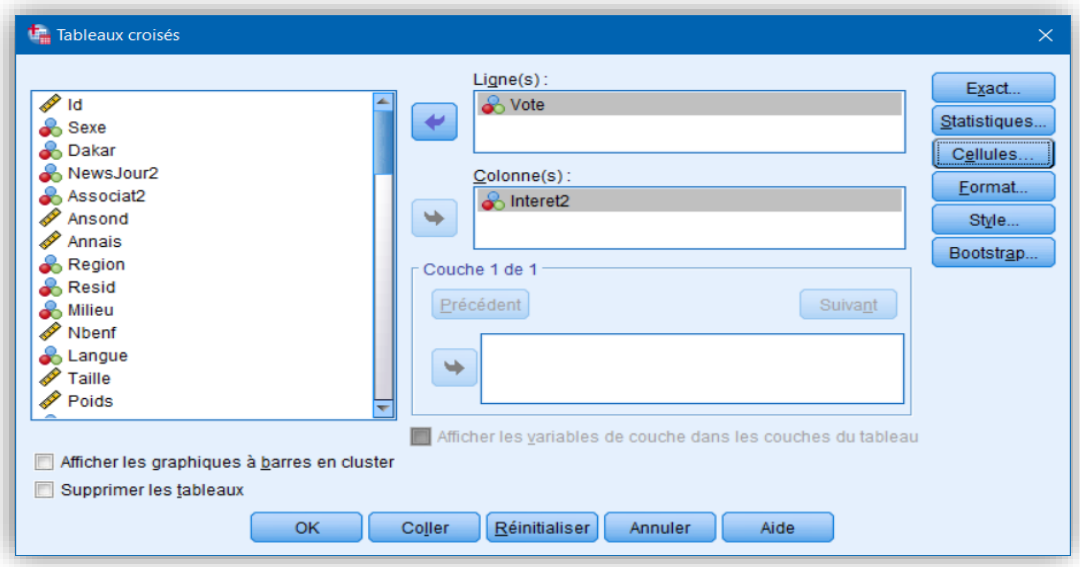
Intérêt pour la politique (X) (Interet2)	→	**Intention de vote (Y) (Vote)**

La question de recherche est la suivante : Y a-t-il une relation d'association entre l'intérêt pour la politique et l'intention de vote chez les étudiants de L2 inscrits en sociologie à l'UGB ? ***Précisément, l'intérêt pour la politique (Interet2) explique-t-il les chances de voter (Vote) ou non chez ces étudiants ?*** On peut penser que plus les gens s'intéressent à la politique, plus ils sont susceptibles de participer aux élections. Pour vérifier cette hypothèse de travail, commençons par construire un tableau croisé bivarié, et calculons le rapport de cotes comme mesure d'association, les deux variables étant catégorielles.

Analyse
Statistiques descriptives
Tableaux croisés



Cliquez sur la commande « **Tableaux croisés** » ! Sélectionnez la VD que vous souhaitez soumettre à l'analyse (vote) et faites-la passer dans le rectangle Lignes(s). Faites de même avec la VI (interet2) de sorte qu'elle se loge, cette fois-ci, dans le rectangle Colonne(s). La capture d'écran ci-dessous en donne l'illustration.



Poursuivez et validez le tout pour obtenir le tableau bivarié en fréquences.

Tableau croisé Vote Seriez-vous oui ou non prêt(e) à voter lors de la prochaine élection présidentielle ? * Interet2 Intérêt pour la politique

Effectif

		Interet2 Intérêt pour la politique		
		,00 Non	1,00 Oui	Total
Vote Seriez-vous oui ou non prêt(e) à voter lors de la prochaine élection présidentielle ?	1,00 Oui	**a** 19	**b** 65	84
	,00 Non	**c** 11	**d** 6	17
Total		30	71	101

1) Calculons la cote de voter chez les étudiants intéressés par la politique (gorupe1)

$$\text{Cote} = \frac{b}{d} = \frac{65}{6} = 10{,}83$$

2) Calculons la cote de voter chez les étudiants non intéressés par la politique (gorupe0)

$$\text{Cote} = \frac{a}{c} = \frac{19}{11} = 1{,}727$$

3) Calculons le rapport de cotes (RC) de voter des étudiants intéressés par la politique (groupe1) par rapport aux étudiants non intéressés par le politique (groupe0)

$$\text{RC} = \frac{b/d}{a/c} = \frac{10{,}833}{1{,}727} = 6{,}27$$

Interprétation statistique : Le rapport de cotes est de 6,27. Les étudiants intéressés par le politique ont 6,27 fois plus de chances de voter comparativement aux étudiants non intéressés par la politique.

4) Calculons le logit de voter des étudiants intéressés par la politique (groupe1) par rapport aux étudiants non intéressés par le politique (groupe0)

$$\text{Logit} = \ln(cote) = \ln(6{,}27) = 1{,}836$$

Si on effectue analyse de régression logistique binaire, le coefficient de régression sera exactement égal à 1,836. La régression logistique approfondit l'analyse du rapport de cote, autant que la régression linéaire approfondit l'analyse de coefficient de corrélation.

2. La régression logistique simple

La VD Vote étant codée 0/1, nous pouvons procéder à l'analyse de régression logistique simple impliquant la VI Interet2.

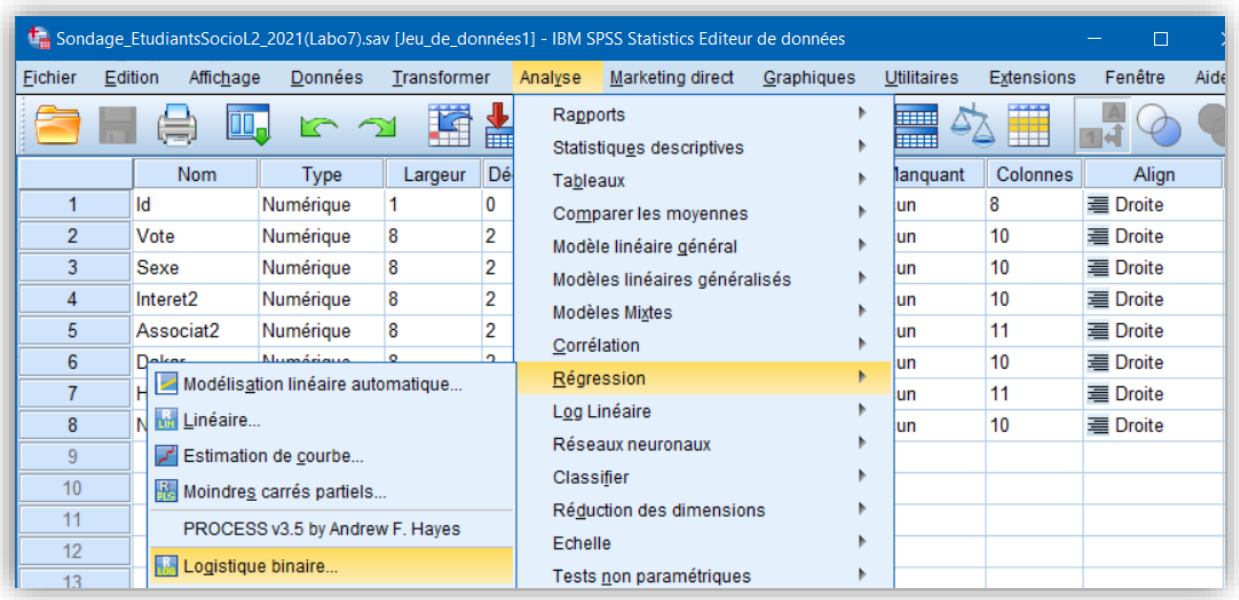
Formulons d'abord les hypothèses statistiques :

H_0 : La probabilité (ou les chances) qu'un étudiant vote est la même, quelque que soit l'intérêt ou non pour la politique (pas de relation dans N).

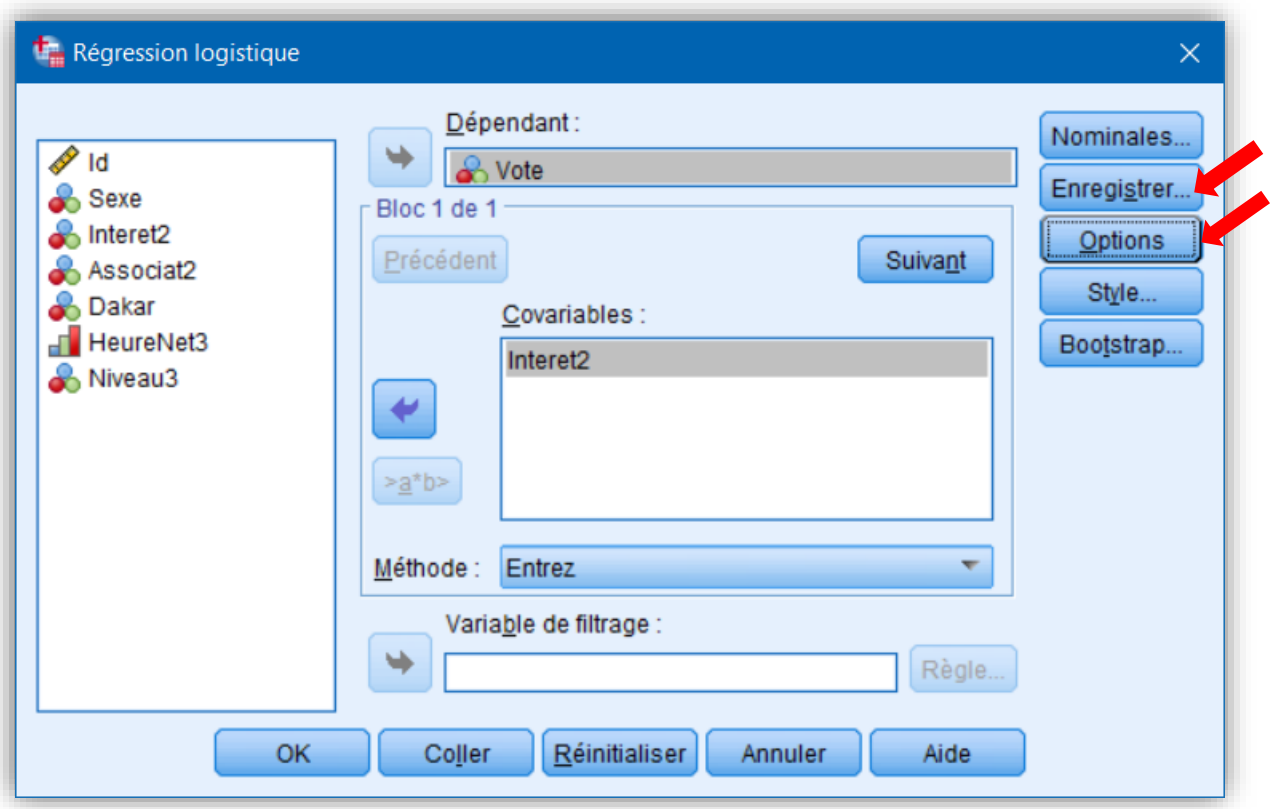
H_1 : La probabilité (ou les chances) qu'un étudiant vote est différente selon l'intérêt ou non pour la politique (relation dans N).

Pour tester l'hypothèse d'une absence de relation et procéder à l'analyse de régression logistique simple, suivons la procédure ci-dessous :

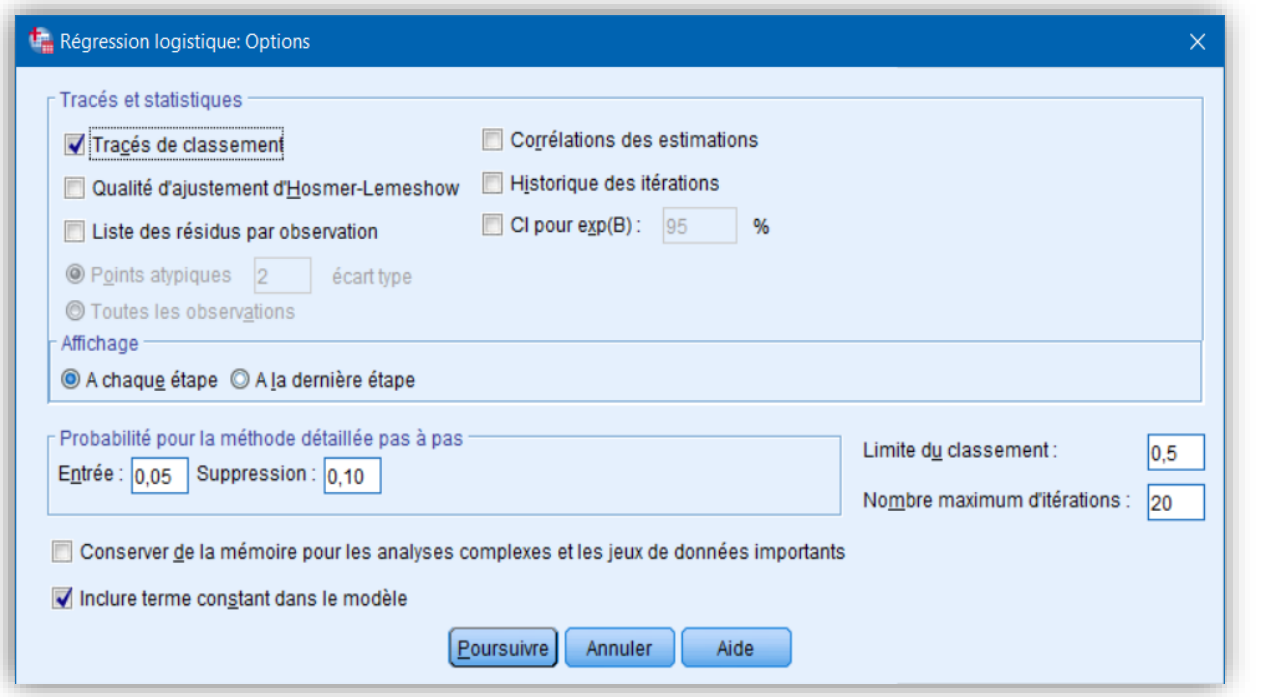
Analyse
Régression
Logistique binaire



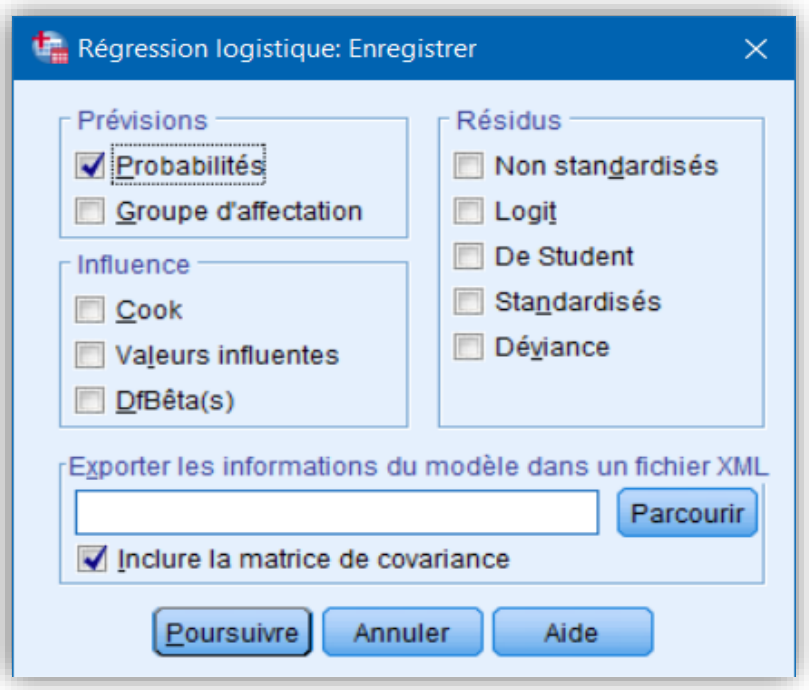
Introduire la variable dépendante sous « Dépendant » (Vote) et la variable indépendante sous « Covariables » (Interet2).



Cliquez sur Options pour sortir les tracés de classement.



Cliquez sur Enregistrer pour cocher probabilités, sous Prévisions.



Poursuivez et validez l'instruction pour obtenir les tableaux de résultats de l'analyse de régression logistique, qui sont nombreux. Par exemple, les trois premiers tableaux indiquent les variables qui ont été introduites dans le modèle d'analyse, et surtout comment la VD et la VI catégorielle ont été codées. Les trois tableaux du *Bloc 0 – Bloc de début* décrivent le modèle nul, soit le modèle sans variable indépendante et qui indique seulement l'ordonnée à l'origine. **Seuls les tableaux «***Variable de l'équation*», « Récapitulatifs des modèles » et « table de classification » du *Bloc 1-Méthode=Introduction* seront interprétés. Ce sont les résultats obtenus à la suite de l'introduction de la VI qui sont les plus informatifs.

1.1. Interprétation du r-deux comme mesure de la force prédictive du modèle

Récapitulatif des modèles

Pas	Log de vraisemblance -2	R-deux de Cox et Snell	R-deux de Nagelkerke
1	80,559[a]	,103	,173

a. L'estimation s'est arrêtée à l'itération numéro 5, car le nombre de modifications des estimations du paramètre est inférieur à ,001.

Le tableau « Récapitulatif des modèles » rapporte, d'une part, la valeur du log de vraisemblance, qui symbolise en quelque sorte la totalité des résidus ou erreurs de prédiction (écarts entre scores réels et scores prédits). Comme en régression linéaire, le logiciel cherche, par itération, à minimiser les résidus. Un meilleur ajustement est obtenu à l'itération numéro 5, soit le log de vraisemblance le plus petit possible.

D'autre part, le tableau précise deux pseudo r-deux : Le R-deux de Cox et Snell et le R-deux de Nagelkerke. Contrairement au R-deux de Cox et Snell, le R-deux de

Nagelkerke peut atteindre le maximum, soit 1. C'est la raison pour laquelle il demeure le coefficient de détermination le plus souvent utilisé. Le R-deux de Nagelkerke s'interpréte de la même façon que le R-deux de la régression linéaire.

Interprétation statistique : L'intérêt pour la politique explique 17% de la variation dans le vote chez les étudiants, le r-deux de Nagelkerke étant de 0,17. La qualité du modèle semble intéressante.

1.2. Interprétation du coefficient de régression logistique comme mesure de l'effet de l'intérêt pour la politique sur le vote

Variables de l'équation

		B	E.S	Wald	ddl	Sig.	Exp(B)
Pas 1[a]	Intérêt pour la politique	1,836	,571	10,354	1	,001	6,272
	Constante	,547	,379	2,081	1	,149	1,727

a. Introduction des variables au pas 1 : Intérêt pour la politique.

Ce tableau, « Variable de l'équation », précise d'abord les coefficients de régression ainsi que leur signification statistique.

a : est la constante (1,836). Soit la valeur du logit de Y lorsque X est nul.

b : coefficient de régression non-standardisé (1,836). Changement dans le logit de la VD (Y) associé à une augmentation d'une unité de la VI (X).

E.S. : Erreur Standard (ou erreur-type);

Wald : test de Wald. Le test de Wald est similaire au test du chi-carré. Pour preuve, au seuil de 0,05 et avec 1 dl, la valeur critique du Wald est de 3,84, soit exactement la même valeur critique du chi-carré au même seuil et au même nombre de degrés de liberté. Le test Wald est au test du chi-carré, ce que le test t est au test z (loi normale). Ce sont deux lois qui suivent une distribution similaire.

ddl: degrés de liberté;

sig: donne la signification statistique du test de Wald pour la constante et la pente. Pour que le test de Wald soit significatif au seuil de 0,05, il faut que sa valeur calculée (10,35 par exemple pour la pente) excède la valeur critique 3,84.

Interprétation statistique : La constante a est de 0,547; ce qui signifie que lorsqu'il y a non-intérêt pour la politique, le logit de l'intention de vote est de 0,546. Le coefficient de régression b est de 1,836, indiquant que l'intérêt pour la politique a un effet positif sur l'intention de vote. Lorsqu'on passe d'une absence d'intérêt à un intérêt pour la politique, le logit (log des chances) de l'intention de vote augmente de 1,836. Le résultat du test de Wald est statistiquement significatif au moins à 99,9% chez l'ensemble des étudiants (Wald=10,35; $p<0,001$).

NB : Il est très difficile d'interpréter le coefficient de régression logistique (b), puisqu'il s'interprète en termes de logit de la VD (Y). Pour faciliter son interprétation en termes

d'intensité, on le transforme en rapport de cote (Exp(b)).

1.3. Interprétation du rapport de cotes comme mesure d'association indiquant la force de la relation entre intérêt pour la politique et intention de vote

Variables de l'équation

		B	E.S	Wald	ddl	Sig.	Exp(B)
Pas 1[a]	Intérêt pour la politique	1,836	,571	10,354	1	,001	6,272
	Constante	,547	,379	2,081	1	,149	1,727

a. Introduction des variables au pas 1 : Intérêt pour la politique.

Le tableau « Variables de l'équation » précise une autre information : le rapport de cotes de la relation Interet2 et Vote.

Exp (B): il s'agit du rapport de cotes (RC), qui est égal à Exp(b) = Exp(1,836) = 6,272. En effet, nous avions appris que cote=Exp(logit).

Quand le rapport de cotes (RC) a une valeur <1, chaque augmentation d'une unité de la VI correspond à une diminution de la chance que l'évènement survienne. Lorsque RC a une valeur >1, chaque augmentation d'une unité de la VI correspond à une augmentation de la chance que l'évènement survienne.

Interprétation statistique : Le rapport de cotes est de 6,272, ce qui signifie que, comparé aux étudiants non intéressés par la politique (0), les étudiants intéressés par la politique ont 6,27 fois plus de chances de voter lors de la prochaine présidentielle. L'effet de l'intérêt pour la politique sur l'intention de vote semble être de grande taille.

1.4. Utilisation de l'équation de régression simple à des fins de prédiction

a) Quelle est la probabilité prédite de voter (Y) si Interet2 (X)=1 ?

Ce genre de question nous amène à calculer la probabilité prédite en fonction des valeurs de la variable indépendante. Utilisons l'équation de régression à cet effet.

$$\text{Logit}(Y) = a + \text{bX}$$

$$\text{Logit}(Y) = 0{,}547 + 1{,}836(X)$$

$$\text{Logit}(Y) = 0{,}547 + 1{,}836(1) = 2{,}383$$

Interprétation statistique : Le logit prédit de l'intention de vote est 2,383 chez les étudiants intéressés par la politique.

Vous aurez compris que les logits sont très difficiles à interpréter et à comprendre. C'est la raison pour laquelle on transforme les logits prédits en probabilités. Autrement dit, pour interpréter les valeurs prédites (logits) en termes exacts de probabilités p, il

faut transformer l'équation de régression logistique. La formule est la suivante :

$$P(Y) = \frac{Exp(a + \mathrm{bX})}{1 + Exp(a + \mathrm{bX})}$$

$$P(Y) = \frac{Exp(2{,}383)}{1 + Exp(2{,}383)} = \frac{10{,}837}{1 + 10{,}837} = 0{,}9155$$

Interprétation statistique : La probabilité qu'un étudiant intéressé par la politique vote est de 91,55%.

b) Quelle est la probabilité prédite de voter (Y) si Interet2 (X)=0 ?

Ce genre de question nous amène à calculer la probabilité prédite en fonction des valeurs de la variable indépendante. Utilisons l'équation de régression à cet effet.

$$\mathrm{Logit}(Y) = 0{,}547 + 1{,}836(0) = 0{,}547$$

Interprétation statistique : Le logit prédit de l'intention de vote est 0,547 chez les étudiants non intéressés par la politique.

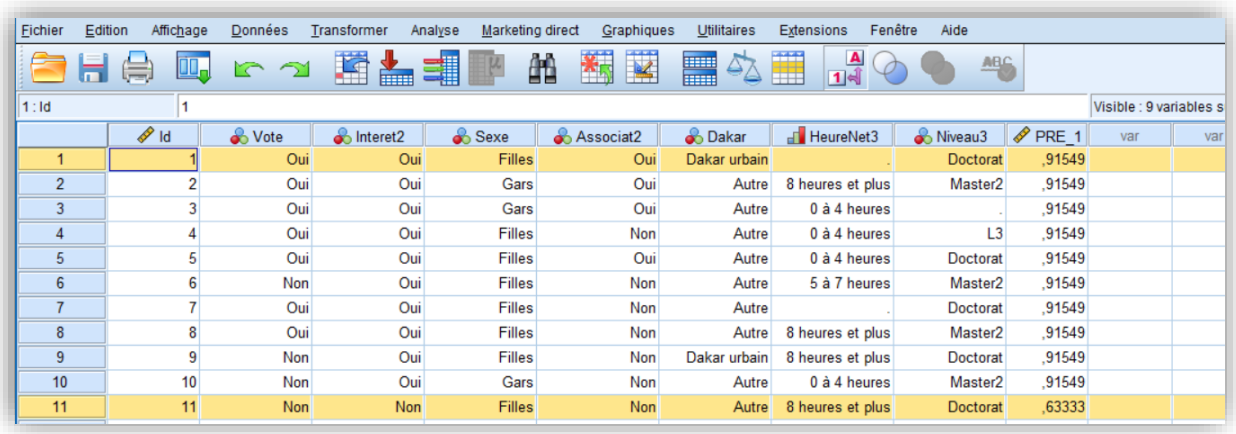
Pour mieux interpréter ce logit prédit, on le transforme en probabilité p. La formule est la suivante :

$$P(Y) = \frac{Exp(a + \mathrm{bX})}{1 + Exp(a + \mathrm{bX})}$$

$$P(Y) = \frac{Exp(0{,}547)}{1 + Exp(0{,}547)} = \frac{1{,}728}{1 + 1{,}728} = 0{,}6333$$

Interprétation statistique : La probabilité qu'un étudiant non intéressé par la politique vote est de 63,33%.

SPSS nous a calculé automatiquement les probabilités prédites et nous les a enregistrées dans une variable (PRE_1).



	Id	Vote	Interet2	Sexe	Associat2	Dakar	HeureNet3	Niveau3	PRE_1	var	var
1	1	Oui	Oui	Filles	Oui	Dakar urbain	.	Doctorat	,91549		
2	2	Oui	Oui	Gars	Oui	Autre	8 heures et plus	Master2	,91549		
3	3	Oui	Oui	Gars	Oui	Autre	0 à 4 heures	.	,91549		
4	4	Oui	Oui	Filles	Non	Autre	0 à 4 heures	L3	,91549		
5	5	Oui	Oui	Filles	Oui	Autre	0 à 4 heures	Doctorat	,91549		
6	6	Non	Oui	Filles	Non	Autre	5 à 7 heures	Master2	,91549		
7	7	Oui	Oui	Filles	Non	Autre	.	Doctorat	,91549		
8	8	Oui	Oui	Filles	Non	Autre	8 heures et plus	Master2	,91549		
9	9	Non	Oui	Filles	Non	Dakar urbain	8 heures et plus	Doctorat	,91549		
10	10	Non	Oui	Gars	Non	Autre	0 à 4 heures	Master2	,91549		
11	11	Non	Non	Filles	Non	Autre	8 heures et plus	Doctorat	,63333		

On retrouve les probabilités prédites qu'on vient de calculer pour les étudiants intéressés par la politique (0,9155) et les pour les étudiants non intéressés par la politique (0,6333). Ce qui est intéressant, c'est que ces probabilités prédites aident à identifier les cas qui sont correctement classés selon leur groupe d'appartenance. Le logiciel s'appuie sur ces informations pour calculer la proportion de cas qui sont correctement classés. Le tableau « table de classification » donne cette proportion.

Tableau de classification[a]

			Prévisions		
			Vote Seriez-vous oui ou non prêt(e) à voter lors de la prochaine élection présidentielle ?		
	Observé		,00 Non	1,00 Oui	Pourcentage correct
Pas 1	Vote Seriez-vous oui ou non prêt(e) à voter lors de la prochaine élection présidentielle ?	,00 Non	0	17	,0
		1,00 Oui	0	84	100,0
	Pourcentage global				83,2

a. La valeur de coupe est ,500

Interprétation statistique : Le taux global de classement est de 83,2%, ce qui signifie que l'intérêt pour la politique prédit correctement le vote 83,2% fois. Autrement dit, 83,2% des cas sont correctement classés dans leur véritable groupe d'appartenance. Encore, nos prédictions sont correctes avec un taux de réussite de 83,2%[1].

1.5. Synthèse : comment présenter les résultats de façon succincte dans un article?

L'analyse de régression logistique montre une relation positive significative entre l'intérêt pour la politique et l'intention de vote chez les étudiants (b=1,836; p<0,001). Comparé aux étudiants non intéressés par la politique, les étudiants intéressés par la politique ont 6,27 fois plus de chances de voter lors de la prochaine présidentielle. Sachant que la probabilité qu'un étudiant intéressé par la politique vote est de 91,55%. Le modèle, plus précisément l'intérêt pour la politique, explique 17% de la variation dans le vote, 83,2% des cas étant correctement classés et prédits.

[1] Par défaut, SPSS retient 0.5 comme point de coupure : il catégorise donc dans les "1" tous les cas qui ont une probabilité plus grande que le point de coupure d'être dans les "1" et dans les "0" tous les cas qui ont une probabilité plus petite que le point de coupure. Souvent, il est nécessaire de changer ce point de coupure lorsque les proportions de la variable dépendante sont déséquilibrées. D'une part, lorsque dans l'échantillon la proportion de l'évènement (catégorie codée 1) est petite (20% par exemple), les probabilités prédites seront petites (autour de 0.20 par exemple), et le point de coupure doit aussi être plus petite (autour de 0.20) pour mieux discriminer les cas. D'autre part, lorsque dans l'échantillon la proportion de l'évènement (catégorie codée 1) est grande (80% par exemple), les probabilités prédites seront grandes (autour de 0.80 par exemple), et le point de coupure doit aussi être plus grande (autour de 0.80) pour mieux discriminer les cas. Pour changer le point de coupure, la procédure SPSS consiste à reprendre la procédure habituelle (menu ANALYSE, puis RÉGRESSION, puis LOGISTIQUE BIBAIRE) et à aller sur OPTIONS pour spécifier la limite du classement.

2. Régression logistique multiple

Pris individuellement, l'intérêt pour la politique explique 17% de la variation dans le vote. Se pourrait-il que d'autres variables comme l'implication dans des associations et le sexe aient un effet discriminant et contribuent à augmenter la proportion expliquée ? *Quel est l'effet de l'intérêt pour la politique sur le vote chez les étudiants, en contrôlant l'effet de l'implication dans les associations et du sexe ? Précisément, quelle est la probabilité (ou les chances) qu'un étudiant intéressé par la politique plutôt que non vote, en contrôlant l'effet de l'implication dans les associations et le sexe ? Encore, quelle est la probabilité (ou les chances) qu'un étudiant impliqué dans les associations plutôt que non vote, en contrôlant l'effet de l'intérêt pour la politique et le sexe ?* Il est possible que l'engagement politique ait un effet plus discriminant que l'engagement social, après avoir contrôlé l'effet du sexe

2.0. Analyses préliminaires

Avant de faire exécuter l'analyse de régression, vérifions la codification des nouvelles variables, de façon à nous assurer qu'elles sont correctement codées. Sortons les fréquences des variables Associat2 et Sexe.

Analyse
Statistiques descriptives
Frequences
- ✓ Associat2 Sexe

Associat2 Implication dans associations

		Fréquence	Pourcentage	Pourcentage valide	Pourcentage cumulé
Valide	,00 Non	48	46,6	47,1	47,1
	1,00 Oui	54	52,4	52,9	100,0
	Total	102	99,0	100,0	
Manquant	Système	1	1,0		
Total		103	100,0		

Sexe Sexe (filles vs gars)

		Fréquence	Pourcentage	Pourcentage valide	Pourcentage cumulé
Valide	,00 Gars	47	45,6	45,6	45,6
	1,00 Filles	56	54,4	54,4	100,0
	Total	103	100,0	100,0	

Les deux prédicteurs étant codées 0/1, on peut les inclure dans le modèle multivarié.

Vérifions s'il n'y pas un problème de multicolinéarité entre les trois prédicteurs.

Analyse

Statistiques descriptives

Frequences

✓ Interet2 Associat2 Sexe

Corrélations

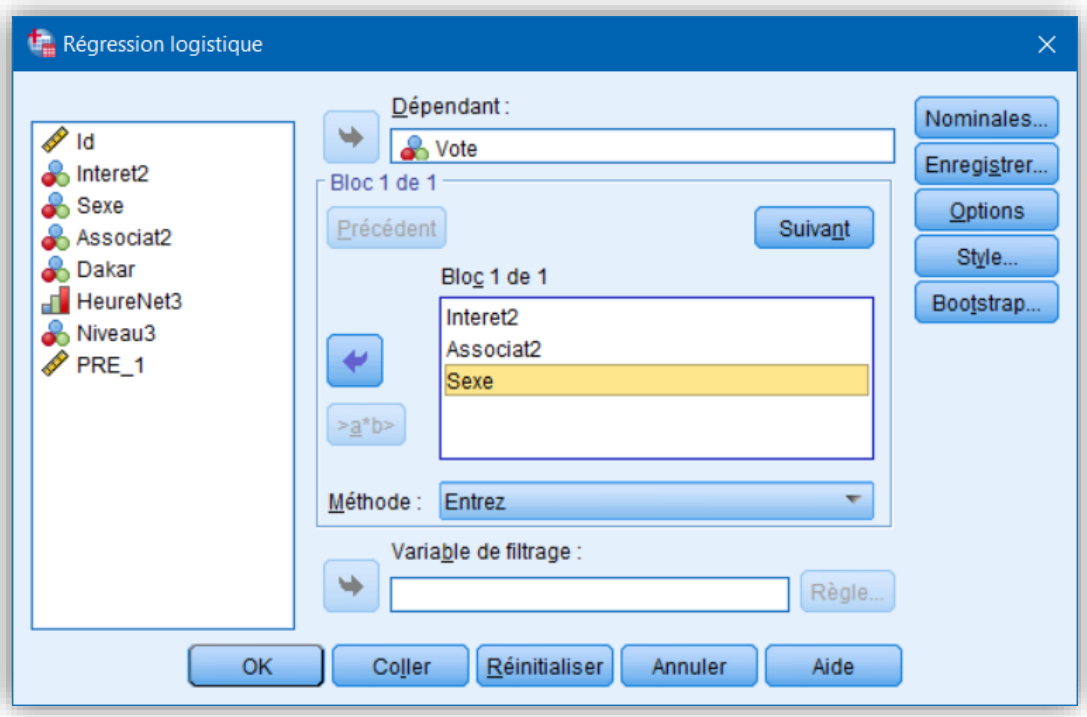
		Interet2 Intérêt pour la politique	Sexe Sexe (gars vs filles)	Associat2 Implication dans associations
Interet2 Intérêt pour la politique	Corrélation de Pearson	1	-,072	-,048
	Sig. (bilatérale)		,472	,635
	N	101	101	100
Sexe Sexe (gars vs filles)	Corrélation de Pearson	-,072	1	-,144
	Sig. (bilatérale)	,472		,149
	N	101	103	102
Associat2 Implication dans associations	Corrélation de Pearson	-,048	-,144	1
	Sig. (bilatérale)	,635	,149	
	N	100	102	102

Il n'y a aucune corrélation supérieure à 0.70. D'ailleurs aucune des corrélations n'est significative statistiquement. Il ne semble pas y avoir de problème de multicolinéarité.

2.1. Procédure pour l'analyse de régression logistique multiple

Pour procéder à l'analyse de régression logistique multiple, suivons la même procédure : **Analyse > Régression > Binaire.**

Dans la fenêtre qui apparaît, toujours dans la boîte « Covariables », rajoutez les variables-contrôle « Associat2 » et « Sexe ».



On garde les mêmes instructions. Validez pour obtenir les résultats ci-dessous :

2.2. Signification statistique du modèle d'ensemble

Tests composites des coefficients du modèle

		Khi-deux	ddl	Sig.
Pas 1	Pas	17,658	3	,001
	Bloc	17,658	3	,001
	Modèle	17,658	3	,001

Interprétation statistique : Ce tableau inique que le modèle d'ensemble (incluant les trois prédicteurs : interet2, associat2, sexe) est statistiquement significatif (p<0,001). Il existe au moins un prédicteur significatif.

2.3. Interpréter les résultats des trois tableaux ci-dessous (exercice)

a) Interprétez le r-deux du modèle (Nagelkerke).

b) Pour chacun des prédicteurs (interet2, associat2, sexe), interprétez le coefficient de régression logistique et sa signification statistique ainsi que le rapport de cote.

c) Interpréter le tableau de classement

Récapitulatif des modèles

Pas	Log de vraisemblance -2	R-deux de Cox et Snell	R-deux de Nagelkerke
1	73,519[a]	,162	,271

a. L'estimation s'est arrêtée à l'itération numéro 5, car le nombre de modifications des estimations du paramètre est inférieur à ,001.

Variables de l'équation

		B	E.S	Wald	ddl	Sig.	Exp(B)
Pas 1[a]	Intérêt pour la politique	2,013	,616	10,682	1	,001	7,483
	Implication dans associations	1,400	,647	4,684	1	,030	4,055
	Sexe (filles vs gars)	-,498	,635	,614	1	,433	,608
	Constante	,132	,691	,037	1	,848	1,142

a. Introduction des variables au pas 1 : Intérêt pour la politique, Implication dans associations, Sexe (filles vs gars).

Tableau de classification[a]

			Prévisions		
			Vote Seriez-vous oui ou non prêt(e) à voter lors de la prochaine élection présidentielle ?		Pourcentage correct
	Observé		,00 Non	1,00 Oui	
Pas 1	Vote Seriez-vous oui ou non prêt(e) à voter lors de la prochaine élection présidentielle ?	,00 Non	4	13	23,5
		1,00 Oui	6	77	92,8
	Pourcentage global				81,0

a. La valeur de coupe est ,500

2.4. Utilisation de l'équation de régression logistique multiple à des fins de prédiction

a) Quelle est la probabilité prédite de voter (Y) si Interet2 (X1)=1, associat2 (X2)=1, Sexe(X3)=1 (fille) ?

Ce genre de question nous amène à calculer la probabilité prédite en fonction des valeurs de la variable indépendante. Utilisons l'équation de régression à cet effet.

$$\text{Logit}(Y) = a + \text{b1X1}+\text{b2X2}+\text{b3X3}$$

$$\text{Logit}(Y) = 0{,}132 + 2{,}013X1 + 1{,}40X2 - 0{,}498X3$$

$$\text{Logit}(Y) = 0{,}132 + 2{,}013(1) + 1{,}40(1) - 0{,}498(1) = 3{,}047$$

Interprétation statistique : Le logit prédit de l'intention de vote est 3,047 chez les étudiants intéressés par la politique (X1=1), impliqué dans des associations (X2=1) et de sexe féminin (X3=1).

Transformons ce logit en probabilité.

$$P(Y) = \frac{Exp(3{,}047)}{1 + Exp(3{,}047)} = \frac{21{,}05}{1 + 21{,}05} = 0{,}9547$$

Interprétation statistique : La probabilité qu'un étudiant intéressé par la politique, impliqué dans des associations et de sexe féminin vote est de 95,47%.

b) Quelle est la probabilité prédite de voter (Y) si Interet2 (X1)=1, associat2 (X2)=1, Sexe(X3)=0 (gars) ?

Ce genre de question nous amène à calculer la probabilité prédite en fonction des valeurs de la variable indépendante. Utilisons l'équation de régression à cet effet.

$$\text{Logit}(Y) = a + \text{b1X1}+\text{b2X2}+\text{b3X3}$$

$$\text{Logit}(Y) = 0{,}132 + 2{,}013X1 + 1{,}40X2 - 0{,}498X3$$

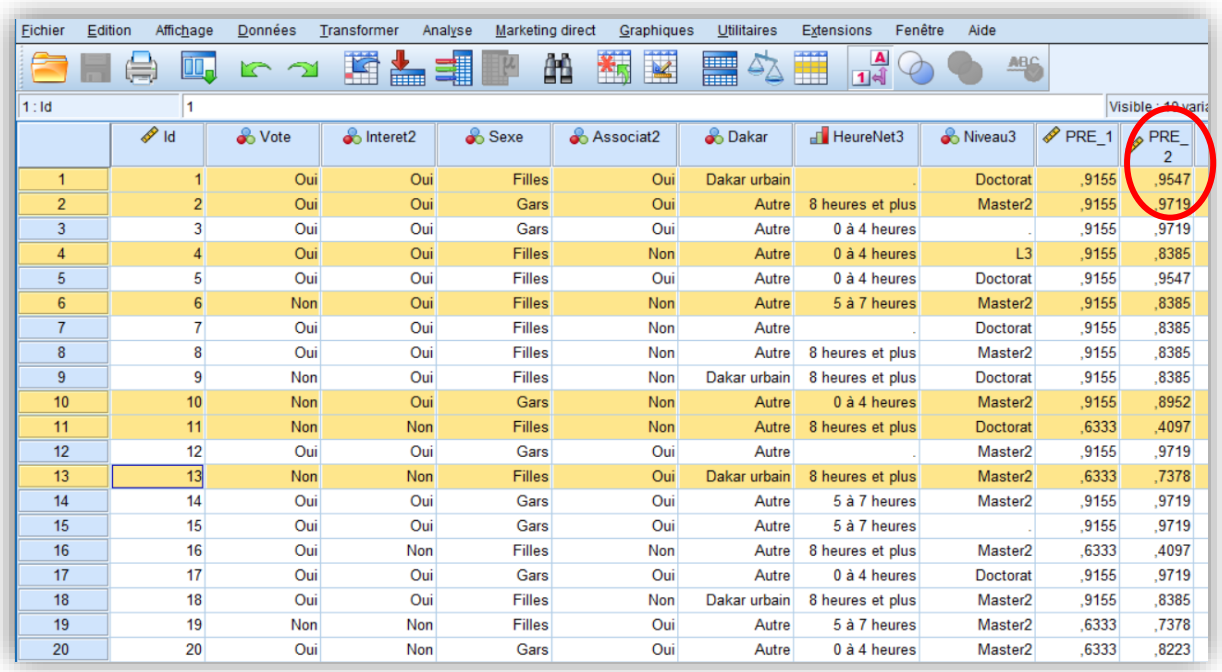
$$\text{Logit}(Y) = 0{,}132 + 2{,}013(1) + 1{,}40(1) - 0{,}498(0) = 3{,}047$$

Interprétation statistique : Le logit prédit de l'intention de vote est 3,545 chez les étudiants intéressés par la politique (X1=1), impliqué dans des associations (X2=1) et de sexe masculin (X3=0).

Transformons ce logit en probabilité.

$$P(Y) = \frac{Exp(3{,}545)}{1 + Exp(3{,}545)} = \frac{34{,}6396}{1 + 34{,}6396} = 0{,}9719$$

Interprétation statistique : La probabilité qu'un étudiant intéressé par la politique, impliqué dans des associations et de sexe masculin vote est de 97,19%.
Ces probabilités prédites calculées sont obtenues à l'aide de SPSS (PRE_2).



Fichier Edition Affichage Données Transformer Analyse Marketing direct Graphiques Utilitaires Extensions Fenêtre Aide

1 : Id | 1 | Visible : 10 varia

	Id	Vote	Interet2	Sexe	Associat2	Dakar	HeureNet3	Niveau3	PRE_1	PRE_2
1	1	Oui	Oui	Filles	Oui	Dakar urbain	.	Doctorat	,9155	,9547
2	2	Oui	Oui	Gars	Oui	Autre	8 heures et plus	Master2	,9155	,9719
3	3	Oui	Oui	Gars	Oui	Autre	0 à 4 heures	.	,9155	,9719
4	4	Oui	Oui	Filles	Non	Autre	0 à 4 heures	L3	,9155	,8385
5	5	Oui	Oui	Filles	Oui	Autre	0 à 4 heures	Doctorat	,9155	,9547
6	6	Non	Oui	Filles	Non	Autre	5 à 7 heures	Master2	,9155	,8385
7	7	Oui	Oui	Filles	Non	Autre	.	Doctorat	,9155	,8385
8	8	Oui	Oui	Filles	Non	Autre	8 heures et plus	Master2	,9155	,8385
9	9	Non	Oui	Filles	Non	Dakar urbain	8 heures et plus	Doctorat	,9155	,8385
10	10	Non	Oui	Gars	Non	Autre	0 à 4 heures	Master2	,9155	,8952
11	11	Non	Non	Filles	Non	Autre	8 heures et plus	Doctorat	,6333	,4097
12	12	Oui	Oui	Gars	Oui	Autre	.	Master2	,9155	,9719
13	13	Non	Non	Filles	Oui	Dakar urbain	8 heures et plus	Master2	,6333	,7378
14	14	Oui	Oui	Gars	Oui	Autre	5 à 7 heures	Master2	,9155	,9719
15	15	Oui	Oui	Gars	Oui	Autre	5 à 7 heures	.	,9155	,9719
16	16	Oui	Non	Filles	Non	Autre	8 heures et plus	Master2	,6333	,4097
17	17	Oui	Oui	Gars	Oui	Autre	0 à 4 heures	Doctorat	,9155	,9719
18	18	Oui	Oui	Filles	Non	Dakar urbain	8 heures et plus	Master2	,9155	,8385
19	19	Non	Non	Filles	Oui	Autre	5 à 7 heures	Master2	,6333	,7378
20	20	Oui	Non	Gars	Oui	Autre	0 à 4 heures	Master2	,6333	,8223

2.5. Présentation et interprétation des résultats

Afin de faciliter l'interprétation des résultats de l'analyse de régression aux fins de diagnostic causal, les résultats sont présentés sous forme tabulaire ou schématique.

Tableau 1. Modèles de régression logistique de l'intention de vote chez les étudiants

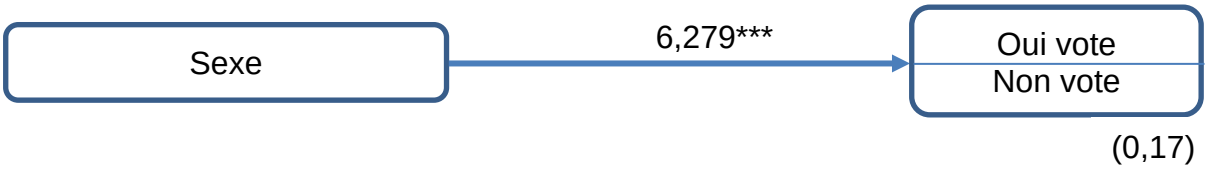
Prédicteurs	**Modèles**	
	1	2
Intérêt pour politique	1,836***	2,01***
	(0,57)	(0,62)
Implication sociale	--	1,40*
		(0,65)
Sexe	--	-0,50
		(0,64)
Constante	0,55	0,13
	(0,38)	(0,69)
R-deux Nagelkerke	0,17	0,27
n	101	100
Taux de classement	0,83	0,81

Notes : Les entrées correspondent à des coefficients de régression logistique multiple (avec les erreurs-types entre parenthèses). *p<0,05 ; ***p<0,001.

NB : Concernant le tableau 1, j'ai décidé de présenter les coefficients de régression logistique multiple. On aurait pu présenter les rapports de cotes comme c'est le cas dans la figure ci-dessous.

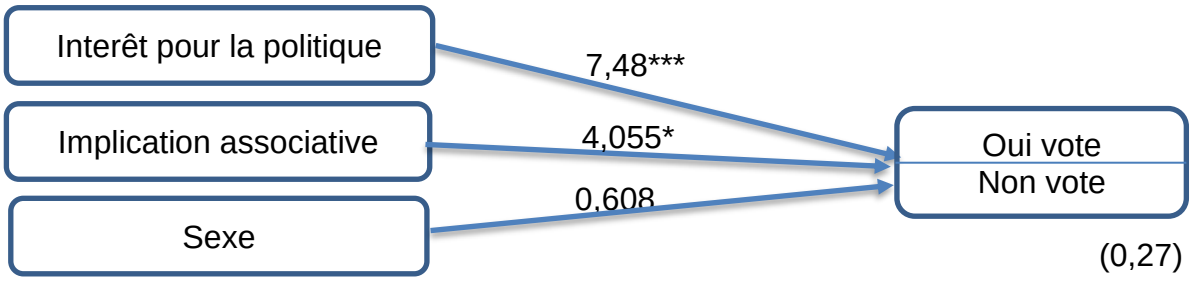
Figure 1. Schématisation des modèles de régression logistique concernant l'intention de vote chez les étudiants

a) Modèle initial



(0,17)

b) Modèle causal complet (discrimination)



(0,27)

Notes : Les chiffres sur les flèches sont des rapports de cotes (1/0,608=1,64, ce qui signifie 1,64 fois moins de chances). Les chiffres sous la variable dépendante correspondent au r-deux de Nagelkerke p<0,05 ; ***p<0,001.

<u>Remarque :</u> Si la régression linéaire multiple met en application le modèle de convergence, la régression logistique multiple met en application plutôt le modèle de la **discrimination**. Mais ces deux modèles sont assez similaires.