Analyse statistique - Séance 4

Aubin Poissonnier-Beraud

Villes et environnements urbains - Université Lumière Lyon 2

2024

Programme de la séance

La régression logistique

Exemple : les expulsions locatives

Le principe de la régression logistique

Nous avons vu jusqu'à présent la modélisation de variables quantitatives. Cependant, il est plus courant en sciences sociales de vouloir étudier des dimensions qualitatives.

Modéliser la probabilité d'un évènement

Les modèles de *régression logistique* sont une extension des modèles linéaires. Plutôt d'estimer une *valeur*, on va tenter d'estimer une *probabilité*. Ils s'appliquent uniquement aux variables catégorielles dichotomiques (deux valeurs possibles).
Les variables peuvent porter sur un phénomène que l'on considère qualitatif « par nature » ; par exemple, il va s'agir d'étudier le passage ou non en classe supérieure, le succès ou l'échec à un concours, le fait d'être propriétaire ou non de son appartement, le fait de voter ou non à une élection, le fait d'avoir un casier judiciaire ou non, etc.
On peut aussi catégoriser et dichotomiser un phénomène quantitatif.

Le logit et les rapports de chances

- On appelle *odd ou chances* le rapport entre la probabilité P d'occurrence de l'évènement binaire considéré et la probabilité $1 - P_i$ de non-occurrence de l'évènement.
  Par exemple, si la probabilité de tomber malade est de $0,8$, la probabilité de ne pas tomber malade est de $1 - 0,8 = 0,2$. L'odd ou les chances valent ainsi $\frac{0,8}{0,2} = 4$: on attend 4 fois plus de personnes malades que de personnes saines.
- On appelle *logit* la fonction qui calcule le logarithme de ces chances.
- On appelle *rapport de chances ou odds ratio* le rapport $\frac{OR_1}{OR_2}$ entre deux chances ou odds.
  Si les chances de tomber malade chez les nourrissons valent 10 contre seulement 5 chez les adultes, l'odd ratio associé vaut $\frac{10}{5} = 2$.
  L'interprétation est complexe car ce n'est pas un simple rapport de probabilités : on ne peut pas dire que la probabilité de tomber malade pour un nourrisson est deux fois plus élevée que celle d'un adulte.

Le logit et les rapports de chances

- On gardera seulement en tête qu'un odd ratio supérieur à 1 signifie que la catégorie au numérateur a de plus grandes chances que la catégorie au dénominateur; on attend donc plus de personnes malades (au lieu de saines) chez les nourrissons que chez les adultes.
- Un odd ratio inférieur à 1 signifie que la catégorie au numérateur a de plus grandes chances que la catégorie au dénominateur; on attend donc moins de personnes malades (au lieu de saines) chez les nourrissons que chez les adultes.

La spécification du modèle

La logique de lecture des modèles logistiques est similaire à celle des modèles linéaires. Il existe néanmoins une difficulté liée à la nature de ce que modèle estime : on ne modélise pas les probabilités en elles-mêmes, mais une fonction qui les contient.

Soient Y la variable dépendante et $X_1, X, ...X_k$ les variables indépendantes,

$$log(\frac{P_i(Y=1)}{1-P_i(Y=1)}) = logit(P_i(Y=1)) = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + ... + \beta_k X_{ki} + \varepsilon_i \quad (1)$$

Retomber sur les probabilités

Sans devoir comprendre le sens mathématique, il faut se souvenir que pour retrouver la probabilité, il faut transformer notre résultat avec la fonction exponentionnelle :

$$P_i = \frac{exp(logit(P_i))}{1 + exp(logit(P_i))} \tag{2}$$

La signification des coefficients

- La lecture d'un modèle logistique **passe toujours par la comparaison avec une catégorie de référence**.
- Le *signe* d'un coefficient (positif ou négatif) nous dit si le passage de la catégorie de référence à telle autre catégorie fait augmenter ou baisser le logit.
- Si le logit augmente, ça veut dire que la probabilité de l'évènement augmente; et inversement.
- Si un logit est plus grand qu'un autre logit, alors l'augmentation de probabilité est plus forte dans un cas que dans l'autre; et inversement.

Exemple : L'enquête PISA

L'enquête PISA est réalisée auprès d'élèves de 15 ans scolarisé.es en France.

L'indice de position sociale

Le statut socio-économique, ou plus précisément le statut économique, social et culturel de l'élève (ESCS) est un indice composite composé à partir de trois indices : le statut socioprofessionnel des parents (basé sur les métiers exercés par les parents), le niveau de formation (nombre d'années d'études) des parents, ainsi que le patrimoine familial. Ce dernier indice inclut lui-même un grand nombre de variables parmi lesquelles les ressources culturelles disponibles dans le foyer (livres), les ressources éducatives (un endroit calme pour travailler, des ouvrages de références ou des logiciels éducatifs) ainsi que d'autres ressources (voiture, connexion internet. . .)

Modéliser la probabilité de redoubler

Modèle reliant le redoublement à la classe d'IPS

On cherche à modéliser la probabilité qu'un élève ait redoublé à partir du quintile d'IPS auquel il appartient. Le premier quintile devient la référence.

$$log(\frac{\text{Probabilité de redoubler}_i}{1 - \text{Probabilité de redoubler}_i}) = logit(\text{Probabilité de redoubler}_i) \quad (3)$$

$$= \beta_0 + \beta_1 \text{Q\_IPS2}_i + \beta_2 \text{Q\_IPS3}_i + \beta_3 \text{Q\_IPS4}_i + \beta_4 \text{Q\_IPS5}_i + \varepsilon_i \quad (4)$$

Table

	Dependent variable:
	IPS
pos_soc_5c2ème	−0.65***
	(0.11)
pos_soc_5c3ème	−1.36***
	(0.13)
pos_soc_5c4ème	−2.01***
	(0.15)
pos_soc_5c5ème	−2.89***
	(0.21)
Constant	−0.55***
	(0.07)
Observations	4,245
Log Likelihood	−1,692.28
Akaike Inf. Crit.	3,394.57
Note:	*p<0.1; **p<0.05; ***p<0.01

Lecture des coefficients

- Le modèle estime que pour la catégorie de référence, un élève du premier quintile d'IPS, le logit du vaut $-0,55$
- Mathématiquement, ça veut dire que $P(\text{redoublement}|IPS = Q1)_i = \frac{exp(-0,55)}{1-exp(0,55)} = 0,37$.
- Passer de la catégorie de référence, le premier quintile d'IPS, au deuxième quintile d'IPS fait baisser de $-0,65$ le logit. Cette élévation sociale fait donc baisser la probabilité du redoublement.
- Mathématiquement, ça veut dire que $P(\text{redoublement}|IPS = Q2)_i = \frac{exp(-0,55-0,65)}{1-exp(-0,55-0,65)} = 0,23$.
- Passer de la catégorie de référence, le premier quintile d'IPS, au troisième quintile d'IPS fait baisser de $-1,36$ le logit. Cette élévation sociale fait donc baisser la probabilité du redoublement.
- Comme $1,36 > 0,55$, **passer dans le troisième quintile réduit plus la probabilité du redoublement que passer dans le deuxième quintile**. On remarque que la probabilité estimée est la plus faible dans le cinquième quintile, soit les élèves issus des classes les plus supérieures.

Modèle reliant le redoublement à la classe d'IPS et au genre

On cherche à modéliser la probabilité qu'un.e élève ait redoublé à partir du quintile d'IPS auquel il appartient et de son genre. Un élève garçon appartenant au premier quintile devient la référence.

$$log(\frac{\text{Probabilité de redoubler}_i}{1 - \text{Probabilité de redoubler}_i}) = logit(\text{Probabilité de redoubler}_i) \quad (5)$$

$$= \beta_0 + \beta_1 \text{Q\_IPS2}_i + \beta_2 \text{Q\_IPS3}_i + \beta_3 \text{Q\_IPS4}_i + \beta_4 \text{Q\_IPS5}_i + \beta_5 \text{Fille} + \varepsilon_i \quad (6)$$

Table

	Dependent variable:
	IPS
pos_soc_5c2ème	−0.66***
	(0.11)
pos_soc_5c3ème	−1.37***
	(0.13)
pos_soc_5c4ème	−2.02***
	(0.15)
pos_soc_5c5ème	−2.90***
	(0.21)
sexeFille	−0.30***
	(0.09)
Constant	−0.40***
	(0.09)
Observations	4,245
Log Likelihood	−1,686.93
Akaike Inf. Crit.	3,385.87
Note:	*p<0.1; **p<0.05; ***p<0.01

Lecture des coefficients

- Le modèle estime que pour la catégorie de référence, un élève **garçon** du premier quintile d'IPS, le logit du redoublement vaut $-0,40$
- Mathématiquement, ça veut dire que $P(\text{redoublement}|IPS = Q1, SEXE = G)_i = \frac{exp(-0,40)}{1-exp(0,40)} = 0,40$.
- Passer de la catégorie de référence, le premier quintile d'IPS, au deuxième quintile d'IPS fait baisser de $-0,66$ le logit. **Indépendamment du genre de l'élève**, cette élévation sociale fait donc baisser la probabilité du redoublement.
- Mathématiquement, ça veut dire que $P(\text{redoublement}|IPS = Q2, SEXE = G)_i = \frac{exp(-0,40-0,66)}{1-exp(-0,55-0,66)} = 0,26$.

Lecture des coefficients

- **Attention**, l'écart mesurée en probabilité $0,40 - 0,26 = 0,12$ ne vaut que pour les garçons !
- Passer de la catégorie de référence, les garçons, aux filles fait baisser de $-0,30$ le logit. **Indépendamment de la position sociale de l'élève**, le fait d'être une fille fait donc baisser la probabilité du redoublement.
- Pour les filles du deuxième quintile, on obtient $P(\text{redoublement}|IPS = Q2, SEXE = F)_i = \frac{exp(-0,40-0,66-0,30)}{1-exp(-0,55-0,66-,30)} = 0,20$.
- L'écart mesuré en probabilité avec le premier quintile d'IPS vaut donc $0,40 - 0,20 = 0,20$.

Conclusion : précaution de lecture

- La lecture en différence de probabilité n'est pas immédiate et porte à se tromper.
- On gardera comme principe de ne commenter que les signes des coefficients ainsi que l'ordre de grandeur (1 coefficient plus grand ou plus petit que l'autre).
- Les coefficients de deux différents modèles logistiques ne doivent pas être comparés en termes de grandeur, par exemple quand on construit deux modèles successifs sur la même population, ou quand on reproduit un même modèle sur des populations différentes.

Qui est un sujet de droit ?

François, Camille. “Déférer au tribunal. Les figures imposées de la défense des locataires au tribunal des expulsions.” Droit et société 106, no. 3 (2020): 527–45."

Question de l'article

"À quelles conditions des ménages populaires et précaires – comme le sont les locataires endetté·e·s en procédure d'expulsion – peuvent-ils constituer des sujets de droit ?" p529

"Dans ce cadre théorique, nous nous demanderons dans quelle mesure les ménages en situation de précarité peuvent être reconnus comme des sujets de droit pleins et entiers, autrement dit comme les sujets effectifs d'un droit générique (tel que le droit civil) qui aille audelà des protections juridiques qui leur sont socialement et spécialement destinées (comme celles du droit social)." p530

Les déterminants des expulsions : tableau croisé

Tableau 1
Présence à l'audience des locataires et décision du ou de la juge

	Expulsion	Délais	Effectifs
Présence à l'audience	24 %	76 %	43 % (266)
Absence à l'audience	65 %	35 %	57 % (356)
Ensemble	48 %	52 %	100 % (622)

Champ : échantillon d'affaires d'expulsion pour dette locative (n = 622, p-value < 0,00001).

Limites des conclusions du tableau croisé

- La différence du taux d'expulsion permet de conclure que la présence réduit les chances de se faire expulser.
- Le lien causal entre la présence et l'expulsion n'est pas assuré : est-ce que les présent.es ne sont simplement pas très différent.es des absent.es ?
- Par exemple, les absent.es seraient des cas perdus d'avance car les dettes sont trop importantes pour qu'un étalement soit envisageable.
- Il serait donc faux de dire que la présence à l'audience est un facteur de protection contre les expulsions.

Les déterminants de la présence à l'audience : modèle de régression

Tableau 2
Déterminants de la présence à l'audience des locataires, régression mixte (n = 622)

		Coeff.	Odds ratios	Effectifs	Présence (moyenne = 43 %)
Constante		-0,58*			
Tribunal situé sur le territoire de la commune du/de la locataire	Non	*Réf.*	1	N = 224 (36 %)	40 %
	Oui	0,48**	1,6	N = 398 (64 %)	45 %
Bailleur	OPHLM	*Réf.*	1	N = 193 (31 %)	54 %
	SA HLM	-0,21	*Ns*	N = 126 (20 %)	48 %
	FTM	-1,14**	3,1^	N = 67 (11 %)	16 %
	SA privée	-0,18	*Ns*	N = 51 (8 %)	41 %
	PP – SCI	-0,74***	2,1^	N = 185 (30 %)	37 %
Connu des services sociaux avant l'assignation	Non	*Réf.*	1	N = 351 (56 %)	40 %
	Oui	0,62***	1,9	N = 172 (28 %)	58 %
	Nr	0,24	*Ns*	N = 99 (16 %)	25 %
Enquête sociale avant audience	Non	*Réf.*	1	N = 485 (78 %)	36 %
	Oui	1,08***	2,9	N = 137 (22 %)	66 %
Statut Matrimonial	Couple	*Réf.*	1	N = 211 (34 %)	50 %
	Homme seul	-0,45**	1,6^	N = 234 (38 %)	30 %
	Femme seule	0,17	*Ns*	N = 177 (28 %)	51 %
Ancienneté dans le logement		1,35. e-03	*Ns*	N = 622 (100 %)	
Dette		6,84.e-05***		N = 622 (100 %)	
Pente de la dette	Diminution	*Réf.*	1	N = 160 (26 %)	53 %
	Augmentation	-0,6***	1,8^	N = 462 (74 %)	39 %

L'utilité du raisonnement toutes choses égales par ailleurs

- La modélisation logistique permet d'interroger l'effet de plusieurs facteurs en réfléchissant comme si le dossier était comparable en tout point, sauf sur le point de la variable considérée
- Par exemple, le montant de la dette augmente plutôt que réduit la probabilité de se présenter à l'audience
- À ce stade, on n'a pas encore modélisé la probabilité de se faire expulser

Le coût judiciaire de la représentation

Tableau 4
Recours à un·e avocat·e des locataires et décision du ou de la juge

	Expulsion	Délais	Effectifs
Tous motifs (n = 790)			
Avocat·e	71 %	29 %	6 % (45)
Sans avocat·e	56 %	44 %	94 % (745)
Total	57 %	43 %	100 % (790)
p-value = 0,052			
Dette locative (n = 622)			
Avocat·e	40 %	60 %	3 % (20)
Sans avocat·e	52 %	48 %	97 % (620)
Total	48 %	52 %	100 % (622)
p-value = 0,49			

Champ : « Base jugements » (n = 790).

Le coût judiciaire de la représentation

		Modèle 1 (AIC = 523,1)	Modèle 2 (AIC = 480,3)	Modèle 3 (AIC = 473,2)	Modèle 4 (AIC = 471,2)	Effectifs (%) (N = 622)
	Constante	-5,25***	-5,85***	-5,06***	-3,73***	
Présence à l'audience	Oui	Ref.	Ref.	Ref.	Ref.	N = 266 (42,8%)
	Non	2,77***	2,69***	2,53***	2,5***	N = 356 (57,2%)
Dette		2,57e-04***	2,53e-04***	2,47e-04***	2,56e-04***	
Pente de la dette	Diminution	Ref.	Ref.	Ref.	Ref.	N = 160 (25,7%)
	Augmentation	1,51***	1,64***	1,71***	1,7***	N = 462 (74,3%)
Opposition du bailleur aux délais	Non	Ref.	Ref.	Ref.	Ref.	N = 470 (75,6%)
	Oui	1,34***	1,85***	1,98***	2,02***	N = 152 (24,4%)
Parc	OPH	Ref.	Ref.	Ref.	Ref.	N = 193 (31%)
	SA HLM	-0,16	-0,32	-0,2	-0,19	N =126 (20,3%)
	FTM	2,00***	2,56***	2,52***	2,56***	N = 67 (10,8%)
	SA Privée	0,37	0,34	0,48	0,47	N = 51 (8,2%)
	Propriétaire individuel – SCI	1,45***	1,48***	1,41***	1,41***	N = 185 (29,8%)
Tribunal	TI 1		Ref.	Ref.	Ref.	N = 234 (37,6%)
	TI 2		-0,37	-0,37	-0,36	N = 172 (27,6%)
	TI 3		2,15***	2,19***	2,3***	N = 115 (18,5%)
	TI 4		0,55	0,56	0,51	N = 101 (16,2%)
Emploi	Non			Ref.	Ref.	N = 49 (7,9%)
	Oui			-1,77**	-1,59**	N = 97 (15,6%)
	nr			-0,67	-0,56	N = 476 (76,5%)
Enfant	Non				Ref.	N = 16 (2,6%)
	Oui				-2,14*	N = 73 (11,7%)
	nr				-1,45 (0,06)	N = 533 (85,7%)

Figure: Déterminants de la probabilité d'expulsion du ménage par le juge