Les modèles de régression comme aide à l'analyse et à la diffusion de résultats dans la statistique publique

Journées suisses de la statistique – Neuchâtel 2016

Nicolas Müller
Responsable de l'information statistique régionale
OCSTAT - Genève

Buts de l'exposé

- Présenter l'utilisation de la modélisation statistique dans deux publications récentes de l'OCSTAT.

- Discuter de la manière de représenter ces résultats dans une publication « grand public ».

Premier exemple : formation continue dans le canton de Genève en 2011

- **Contexte** : publication sur la formation continue dans le canton de Genève
- **Source** : microrecensement formation de base et continue (MRF) mené par l'OFS, densification de l'échantillon pour Genève (2 000 résidants interrogés)
- **Méthode** : régression logistique
- **Variable expliquée** : participation à une formation contenue au cours de l'année écoulée (oui/non)
- **Variables explicatives** : profil socio-démographique et professionnel

Régression logistique : exemple avec une variable explicative

- Le logarithme de la cote des probabilités est expliqué par une variable catégorielle X avec *i* modalités représentée par *i-1* variables indicatrices :

$$ln\ \frac{p(Y=1\mid X)}{1-p(Y=1\mid X)} = \beta_0 + \beta_1 x_1 + \ \ldots\ + \beta_{i-1} x_{i-1}$$

  - Exemple : Y représente le fait de suivre une formation continue (y=1) ou non (y=0), X représente le sexe de l'individu (x_1=1 pour les femmes, x_1=0 pour les hommes).

- On peut représenter les résultats de deux façons principales :

  - L'exponentielle du coefficient β_1 représente le **rapport des cotes** (odds ratio) de la modalité 1 de la variable *x* et de la modalité de référence 0 :

$$e^{\beta_1} = \frac{p(Y = 1 \mid X = 1)}{1 - p(Y = 1 \mid X = 1)} \Big/ \frac{p(Y = 1 \mid X = 0)}{1 - p(Y = 1 \mid X = 0)}$$

    - Interprétation : si e^{β_1} = 2, alors les chances de suivre une formation continue (cote p/(1-p)) sont deux fois plus élevées parmi les femmes que parmi les hommes.

  - La **probabilité que Y soit égale à 1** si X=1 peut être calculée :

$$p(Y = 1 \mid X = 1) = \frac{e^{\beta_0 + \beta_1}}{1 + e^{\beta_0 + \beta_1}}$$

    - Interprétation : la probabilité de suivre une formation continue lorsqu'on est une femme est de 0,75 (la probabilité pour un homme est de 0,6, loin de 2 fois moins !).

Utilité du modèle multivarié

- En ajoutant une variable explicative, les effets de chaque variable :

$$ln\,\frac{p(Y=1\mid X,Z)}{1-p(Y=1\mid X,Z)} = \beta_0 + \beta_1 x_1 + \beta_2 z_1$$

- e^{β_1} est le rapport de cote entre les femmes et les hommes, quel que soit la valeur de Z ("et ceteris paribus")

$$e^{\beta_1} = \frac{p(Y=1\mid X=1,Z)}{1-p(Y=1\mid X=1,Z)} \Big/ \frac{p(Y=1\mid X=0,Z)}{1-p(Y=1\mid X=0,Z)}$$

Utilité du modèle multivarié (suite)

- L'introduction successive de variables explicatives dans le modèle permet de tester si l'effet d'une variable est "fallacieux" ou non :
  - Exemple : la probabilité de suivre une formation continue est-elle plus élevée chez les hommes que chez les femmes, même à niveau de revenu équivalent ? Ou l'effet du sexe relaie-t-il un effet du revenu ?
- Sans multiplier les tableaux croisés descriptifs, la modélisation permet d'identifier plus rapidement ce genre d'effet.

Utile pour le statisticien, mais pour le lecteur ?

- La signification des coefficients d'une régression logistique (rapports de cote) n'est pas immédiatement compréhensible.

- Publier un tableau avec la probabilité pour chaque profil de suivre une formation continue semble indigeste.

- Cependant, un modèle bien ajusté permet de résumer un processus de réflexion (sélection des variables, choix du meilleur modèle).

Solution adoptée : pas de chiffres, mais des couleurs

Facteurs favorisant la participation à la formation continue parmi les actifs de 25 - 64 ans

Facteurs	Influence
Sexe	
Homme	*référence*
Femme	+
Légende :	
augmente fortement la probabilité de participation	+++
	++
	+
peu ou pas différent de la catégorie de référence	+ / -
	-
	- -
diminue fortement la probabilité de participation	- - -

A profil équivalent (pour toutes les autres variables) :

- Une femme a plus de chance de participer à une formation continue qu'un homme;
- Une personne âgée de 45 ans ou plus a moins de chance qu'une personne âgée de 25-34 ans de participer à une formation continue.

Facteurs favorisant la participation à la formation continue parmi les actifs de 25 - 64 ans

Facteurs	Influence
Sexe	
Homme	*référence*
Femme	+
Age	
25-34 ans	*référence*
35-44 ans	+ / -
45-54 ans	-
55-64 ans	- -
Origine	
Suisse	*référence*
Etranger	-
Revenu annuel brut	
40 000 francs ou moins	*référence*
De 40 001 à 80 000 francs	+
De 80 001 à 120 000 francs	++
Plus de 120 000 francs	+++
Niveau de formation	
Degré secondaire 1	*référence*
Degré secondaire 2	+++
Degré tertiaire	+++
Catégorie de profession	
Ouvriers, artisans, conducteurs	*référence*
Professions intermédiaires et des services	+++
Professions intellectuelles, scientifiques et de direction	+++
Statut professionnel	
Salarié	*référence*
Non salarié	-
Fonction d'encadrement	
Sans fonction d'encadrement	*référence*
Avec fonctions d'encadrement	+
Taille de l'entreprise	
1 à 19 employés	*référence*
20 à 199 employés	+
200 employés ou plus	+

Deuxième exemple : facteurs explicatifs des loyers

- **Contexte** : publication sur les conditions d'habitation dans le canton de Genève
- **Source** : relevé structurel, données poolées 2011-2013 (47 030 ménages)
- **Méthode** : régression linéaire
- **Variable expliquée** : logarithme du loyer (car distribution très étalée à droite)
- **Variables explicatives** : profil du ménage, caractéristiques du logement et du bâtiment

Modèle retenu

- ln(loyer) = constante + type de ménage + âge du ménage + commune de résidence + niveau de formation + nombre de pièces du logement + année de construction

- Toutes les variables sont catégorielles et codées avec des variables indicatrices (dummies)

- En raison de la forme du modèle (transformation logarithmique), les coefficients correspondent à un multiplicateur de loyer

Présentation des résultats

- Le loyer moyen pour un profil de référence est une valeur facilement interprétable par les lecteurs :

  – Un couple avec enfant, dont l'âge du conjoint le plus âgé se situe entre 35 et 45 ans, avec une formation de degré tertiaire, habitant un logement de trois pièces construit entre 1971 et 1980 et se situant à Genève paie en moyenne 1543 CHF.

- L'exponentielle des coefficients représente le multiplicateur de loyer par rapport à la catégorie de référence : représenté en %, il est facilement interprétable.

Facteurs d'influence du loyer (1/2) :

Profil de référence :

- Couple avec enfant(s)
- De 35 à moins de 45 ans
- Ville de Genève
- Formation de degré tertiaire
- Logement de 3 pièces
- Bâtiment construit en 1971 et 1980

Facteurs	Catégories	Influence en % par rapport à la catégorie de référence
Loyer du profil de référence : 1543.-		
Type de ménage	Couples avec enfant(s)	référence
	Couples sans enfant	2,0
	Ménages d'une personne	1,4
	Pères et mères seuls avec enfant(s)	-1,6
	Ménages non familiaux	( )
	Ménages multifamiliaux	-3,9
Age du ménage	Moins de 25 ans	( )
	De 25 à moins de 35 ans	3,6
	De 35 à moins de 45 ans	référence
	De 45 à moins de 55 ans	-6,0
	De 55 à moins de 65 ans	-10,0
	De 65 à moins de 75 ans	-18,2
	75 ans ou plus	-20,1
Commune de résidence	Genève	référence
	Bernex	( )
	Carouge	-5,7
	Chêne-Bougeries	( )
	Chêne-Bourg	( )
	Collonge-Bellerive	8,9
	Grand-Saconnex	-3,2
	Lancy	-2,8
	Meyrin	-2,7
	Onex	-7,1
	Plan-les-Ouates	-3,0
	Thônex	-6,9
	Vernier	-6,2
	Versoix	-3,8
	Veyrier	6,8
	Autres communes du canton de Genève	3,9

Facteurs d'influence du loyer (2/2) :

Niveau de formation	Degré tertiaire	référence
	Degré secondaire 2	-12,3
	Degré secondaire 1	-16,0
Nombre de pièces habitables du logement	1	-40,4
	2	-20,0
	3	référence
	4	18,2
	5	38,9
	6	87,4
	7 ou plus	224,6
Année de construction du bâtiment	Avant 1945	( )
	De 1946 à 1960	-8,5
	De 1961 à 1970	-8,6
	De 1971 à 1980	référence
	De 1981 à 1990	7,4
	De 1991 à 2000	4,9
	Après 2000	8,1

Conclusion… et discussion

- Attitude orientée utilisateurs (la complexité lui est cachée).
- Répond aux exigences de publications plus courtes.
- Impact médiatique souvent plus fort pour des résultats très synthétiques que pour des analyses approfondies.