Modèles de comptage appliqués aux décisions de candidature aux offres d'emploi sur le web

Julie Séguéla et Gilbert Saporta

Laboratoire Cédric - CNAM, 292 rue Saint Martin, 75141 Paris cedex 03, France

Abstract

Count data regression models are widely used technics in the literature. We propose to applicate three types of Poisson models to predict job offers applications online in a two-step way (to alert the recruiter in case of very low return, or to predict the count in the other case). In addition to the commonly used Poisson regression model, we carry out hurdle and zero-inflated regressions, two-part models which address the issue of excess zeroes and overdispersion. Contrary to the standard Poisson regression, hurdle and zero-inflated predict alerts efficiently but are less satisfying in the evaluation of application counts.

Résumé

Les modèles pour données de comptage sont des techniques largement étudiées dans la littérature. Nous proposons une application de trois types de modèles poissonniens pour modéliser les candidatures aux offres d'emploi sur le web en deux temps : alerte en cas de retour très faible, sinon estimation de l'effectif. Pour cela, en plus de la régression de Poisson standard, nous mettons en œuvre des modèles construits en deux étapes : hurdle et zero-inflated, adaptés pour la modélisation des zéros en excès et les problèmes de sur-dispersion. Efficaces pour la prédiction des alertes, ces méthodes s'avèrent moins performantes pour la modélisation de l'effectif, à l'inverse de la régression de Poisson standard.

Mots-clés : données de comptage, Poisson, hurdle, zero-inflated, offres d'emploi

1. Introduction

La modélisation de données de comptage est une problématique très répandue dans divers domaines comme la banque, les assurances, l'économétrie, la médecine ou encore le marketing. Aussi, les méthodes de modélisation adaptées à ce type de données ont été largement explorées dans la littérature. La régression de Poisson est le recours standard dans ce genre de situation, cependant, de nombreuses applications à des cas réels ont mis en évidence la nécessité de trouver des solutions alternatives permettant de gérer les problèmes sur-dispersion et les excès de zéros induits par les mécanismes du phénomène étudié. Parmi les alternatives existantes, les régressions hurdle (Mullahy, 1986) et zero-inflated (Lambert, 1992) répondent de manière spécifique au problème des zéros en excès tout en gérant la sur-dispersion des données. Les travaux de recherche sur la généralisation de ces modèles ainsi que leurs mises en application sont nombreux. Consul et Famoye (1992) proposent une régression de Poisson généralisée avec l'introduction d'un nouveau paramètre dans le modèle standard pour modéliser la dispersion. Récemment, Famoye et Singh (2006) développent une régression de Poisson généralisée zero-inflated pour modéliser les violences domestiques.
Dans cette étude, nous mettons en application des modèles pour données de comptages dans le but d'expliquer les décisions de candidatures aux offres d'emploi sur le web. Multiposting.fr [1],

1. http ://www.multiposting.fr/

outil de multidiffusion d'offres d'emploi sur Internet, met à disposition pour cette étude une base de données d'offres d'emploi ayant été postées sur une centaine de sites d'emploi entre décembre 2008 et janvier 2010. Nous nous concentrons sur l'analyse des offres postées sur un site généraliste pour lesquelles nous connaissons le nombre total d'affichages du contenu de l'annonce (clic sur un lien) et le nombre final de candidatures effectives (envoi d'un CV) à l'issue de la campagne. Pour avoir accès au texte complet de l'offre, les candidats potentiels doivent cliquer sur le lien défini par le titre du poste (un bref extrait de l'offre est présenté sous ce lien), ils ont donc déjà pris connaissance de certaines informations. Nous allons modéliser le nombre de CV reçus connaissant le nombre d'affichages de l'offre, ce qui revient à modéliser le taux de "transformation" des candidatures potentielles en candidatures réelles. Une deuxième problématique consistera à détecter les offres engendrant un nombre de candidatures très faible voire nul (un seuil sera fixé) afin de pouvoir anticiper ces situations.
Dans la section 2, nous décrivons les données relatives aux offres étudiées. Les modèles mis en œuvre sont présentés dans la section 3. Ensuite, nous comparons les qualités d'ajustement des modèles ainsi que leurs capacités de généralisation (cf. section 4). Enfin, la section 5 est consacrée aux conclusions et perspectives de recherches pour l'amélioration du modèle.

2. Présentation des données

Nous étudions des offres d'emploi ayant été diffusées sur un site généraliste via la plateforme de multidiffusion Multiposting.fr. Une des spécificités de cette solution étant de répondre aux difficultés de recrutement des entreprises, les données comprennent une part importante de postes faisant référence à des profils rares. Par conséquent, une proportion importante des offres est associée à un nombre de candidatures très faible voire nul. Nous restreignons le périmètre étudié aux offres ayant généré au moins un affichage, le nombre de candidatures (CV) étant trivialement nul dans le cas contraire. Suite à ce filtrage, nous avons 175 observations.

Variable	Libellé	Moyenne (écart-type)
EXPE_15	1 à 5 ans d'expérience requise	0.71
REG_P	Poste à pourvoir dans la région parisienne	0.37
REG_NE	Poste à pourvoir dans la région nord-est	0.27
FCT_IND	Domaine d'activité : industrie	0.18
FCT_BTP	Domaine d'activité : BTP	0.06
FCT_COM	Domaine d'activité : commercial, marketing	0.29
FCT_GEST	Domaine d'activité : gestion, finance, banque	0.13
FCT_INFO	Domaine d'activité : informatique	0.13
FCT_RH	Domaine d'activité : ressources humaines	0.07
NB_CLIC	Nombre d'affichages de l'offre	152.3 (159.5)
NB_CV	Nombre de candidatures effectives	8.0 (14.3)
NB_CV/NB_CLIC	Taux de transformation des affichages en candidatures	4.1% (3.3%)

Tableau 1 : Les variables du modèle

Pour caractériser les offres, nous disposons de données catégorielles définies par la nomenclature du site d'emploi : le type de contrat, le niveau d'études requis, le niveau d'expérience requis, la zone géographique où le poste est à pourvoir, le secteur et le domaine d'activité du poste. Nous connaissons également le nom du diffuseur et le nombre d'affichages de l'offre, en d'autres termes, le nombre total de candidatures potentielles. Des premières expérimentations nous

conduisent à retenir un ensemble de variables explicatives pour le modèle (cf. tableau 1). Les variables EXPE_15 à FCT_RH sont dichotomiques.
La figure 1 présente les fréquences observées du nombre de candidatures effectives (zoom sur les valeurs entre 0 et 40 CV) et la distribution théorique d'une loi de Poisson de paramètre égal à la moyenne empirique. Il apparaît clairement qu'une telle distribution surestime le nombre de valeurs moyennes et sous-estime le nombre de valeurs faibles. Nous avons également affaire à un problème de sur-dispersion.

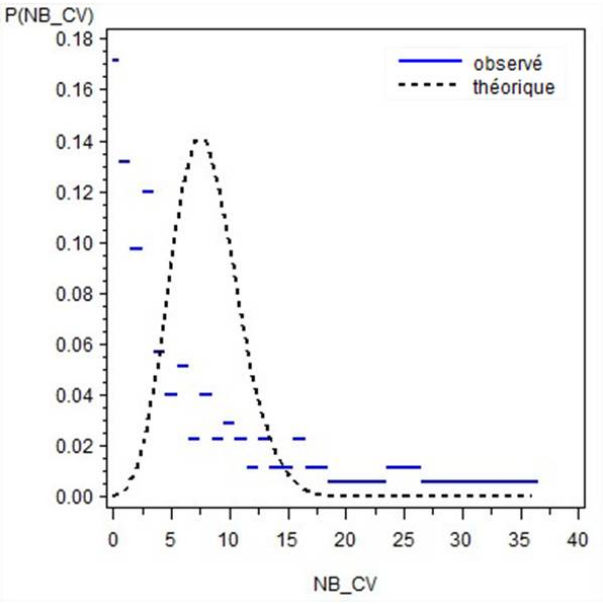

Figure 1 : Distribution empirique de NB_CV et approximation par une loi de Poisson

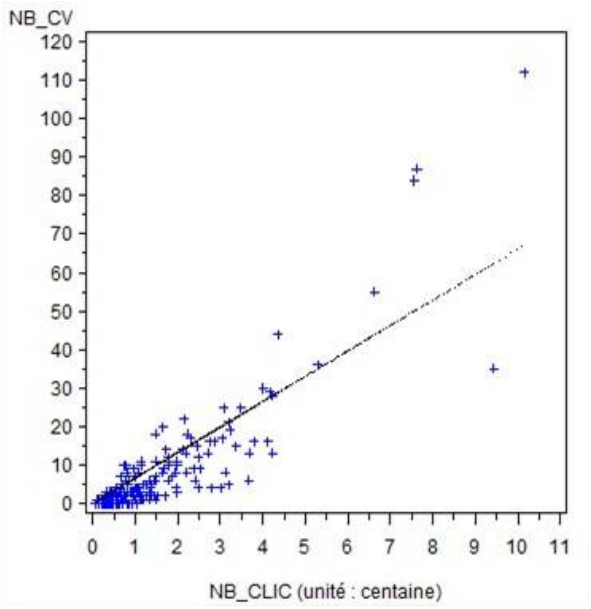

Figure 2 : NB_CV en fonction de NB_CLIC et droite de régression sans constante

La figure 2 représente le nuage de points du nombre de candidatures en fonction du nombre d'affichages, ainsi que la droite de régression associée (constante mise à 0). Le nuage de points suggère une relation non linéaire entre le nombre de CV reçus et le nombre d'affichages de l'offre. La variable NB_CLIC sera donc introduite dans le modèle en tant que terme de calage (paramètre "offset") et en tant que variable explicative.

3. Modélisation

Afin d'alerter les recruteurs lorsque le risque de recevoir un nombre insuffisant de candidatures est élevé, nous modélisons un processus en deux étapes :

- Prédire si le nombre de CV est inférieur ou égal à 3 (seuil choisi arbitrairement d'après l'observation de la figure 1), ou s'il est supérieur ;
- S'il est supérieur, estimer le nombre de candidatures reçues.

En d'autres termes, la prédiction exacte du nombre de CV ne nous intéresse pas si celle-ci est inférieure ou égale à 3. Afin de répondre à cette problématique, nous allons tester trois types de modèles. D'abord, la régression de Poisson avec paramètre de dispersion, puis deux variantes seront testées : les régressions hurdle et zero-inflated. Pour la suite, soient Y_i la variable réponse observée, N_i le terme de calage (ici le nombre d'affichages), X_i le vecteur des variables explicatives et β le vecteur des coefficients du modèle. n et p désignent respectivement le nombre d'observations et le nombre de paramètres du modèle.

3.1. Régression de Poisson

Nous supposons : $Y_i|X_i \sim P(\mu_i)$ avec $log(\mu_i) = X_i\beta + log(N_i)$. Nos premières expérimentations mettent en évidence un problème de sur-dispersion (variance supérieure à la moyenne). Nous introduisons donc un paramètre de dispersion ϕ pour relâcher les hypothèses de restriction sur la variance : $V(Y_i|X_i) = \phi\mu_i$. Bien qu'il n'affecte pas l'estimation des coefficients du modèle, le paramètre ϕ va fournir un terme de correction rendant exploitables les tests d'inférence statistique. Comme le suggèrent McCullagh et Nelder (1989), nous estimons ϕ par le ratio entre le Chi-deux de Pearson et son degré de liberté associé : $\phi = \frac{\sum_{i=1}^{n} \frac{(Y_i-\mu_i)^2}{V(\mu_i)}}{n-p}$.

3.2. Régression hurdle

La régression hurdle (Mullahy, 1986), qui correspond à un modèle en deux étapes, s'adapte parfaitement à notre problème. Nous étudions sa forme la plus populaire : le modèle Logit-Poisson. Dans un premier temps, une régression logistique permet de déterminer si l'effectif est nul ou non nul, puis dans un deuxième temps l'effectif est modélisé par une régression de Poisson tronquée. Nous utilisons les mêmes variables explicatives pour les deux modèles, le vecteur des coefficients de la régression logistique est nommé α. Adaptée à notre problème, la fonction de densité est :

$$f(Y_i \mid X_i) = \begin{cases} P(Y_i \leq 3) = \frac{e^{X_i\alpha}}{1+e^{X_i\alpha}} & , Y_i \leq 3 \\ [1 - P(Y_i \leq 3)] \frac{e^{-\mu_i}\mu_i^{Y_i}}{\left[1-e^{-\mu_i}(1+\mu_i+\frac{1}{2}\mu_i^2+\frac{1}{6}\mu_i^3)\right]Y_i!} & , Y_i > 3 \end{cases} \text{ avec } log(\mu_i) = X_i\beta + log(N_i).$$

3.3. Régression de Poisson zero-inflated

Introduite par Lambert (1992), la régression zero-inflated est une autre alternative en cas de sur-dispersion faisant appel à une modélisation en deux étapes. Elle diffère de la régression hurdle dans le sens où les effectifs nuls peuvent être générés par les deux processus. D'abord, un modèle logit détermine si l'observation provient du groupe où le résultat est toujours nul ou s'il elle provient du groupe où les valeurs peuvent être positives ou nulles. Ensuite, une régression de Poisson estime les effectifs pour ce deuxième groupe. En l'adaptant à notre problème, la fonction de densité est :

$$f(Y_i \mid X_i) = \begin{cases} p_i + (1-p_i)\frac{e^{-\mu_i}\mu_i^{Y_i}}{Y_i!} & , Y_i \leq 3 \\ (1-p_i)\frac{e^{-\mu_i}\mu_i^{Y_i}}{Y_i!} & , Y_i > 3 \end{cases} \text{ où } log(\mu_i) = X_i\beta + log(N_i), p_i = \frac{e^{X_i\alpha}}{1+e^{X_i\alpha}}$$

et $(1-p_i)$ est la probabilité que Y_i suive une loi de Poisson de paramètre μ_i.

Pour la mise en application des modèles, nous utilisons le logiciel SAS et la procédure NLMIXED qui permet de spécifier les équations du modèle avec une estimation par maximum de vraisemblance (cf. Liu et Cela, 2008). Pour évaluer les capacités de prédiction des différents types de modèles, nous calculons les erreurs de prédiction par validation croisée [2]. Aussi, nous choisissons de retenir des modèles plus parcimonieux en retirant les facteurs non significatifs à 5% par une méthode de type descendante.

4. Résultats

Les estimations obtenues pour les trois modèles mis en œuvre sont présentées dans le tableau 2.

2. L'échantillon des données est divisé en 10 parties. Sur chaque sous-échantillon, nous évaluons les erreurs de prédiction à partir du modèle estimé sur l'ensemble des données formé par les 9 autres sous-échantillons.

	Régression de Poisson	Régression hurdle		Régression zero-inflated	
		Logit	Poisson	Logit	Poisson
INTERCEPT	-4.4420**	6.5776**	-4.3055**	6.0902**	-4.1867**
EXPE_15	0.2816**	-0.0967	0.3296**	-0.3080	0.3262**
REG_P	0.7019**	-2.0436**	0.7525**	-1.7729*	0.7037**
REG_NE	0.2593*	-1.7631*	0.2429	-1.1602	0.2726*
FCT_IND	0.6984**	-1.9300*	0.5842**	-2.1068	0.5618**
FCT_BTP	0.5817**	3.5123	0.9906**	3.4472	0.8904**
FCT_COM	0.6431**	-0.6380	0.5625**	-0.9812	0.5061**
FCT_GEST	0.5370**	-0.9339	0.4852**	-1.1858	0.4630**
FCT_INFO	0.5002**	-0.3718	0.4165*	-0.9995	0.3698
FCT_RH	0.6330**	-2.2700*	0.4164*	-4.8569	0.3943*
NB_CLIC	0.683×10^{-3}**	-0.03766**	0.460×10^{-3}**	-0.03625**	0.408×10^{-3}**
log(NB_CLIC)	1		1		1
Nb param.	11		22		22
Log-vrais.	-411,4		-280.5		-288.1
AIC	844.8		605.1		620.2
BIC	879.6		674.8		689.8

paramètre significatif à : 5% (*), 1% (**)

Tableau 2 : Coefficients estimés pour les trois modèles

La statistique de test du rapport de vraisemblance comparant le modèle hurdle au Poisson $-2(-411.4+280.5)$ indique un large rejet du modèle de Poisson, également rejeté contre le modèle zero-inflated. Les critères de choix de modèle AIC et BIC viennent appuyer ces conclusions malgré l'augmentation importante du nombre de paramètres.
Nous souhaitons maintenant comparer les capacités de prédiction des différents modèles. Le tableau 3 présente les erreurs de classement associées à la prédiction de l'événement "moins de 3 CV reçus" (nous l'appelons E). Pour les régressions hurdle et zero-inflated, il est nécessaire de spécifier un seuil pour la probabilité estimée par le modèle logit au-delà duquel nous prédisons E. Après plusieurs expérimentations, nous choisissons de présenter les résultats pour trois valeurs du seuil.

Modèle	Seuil	$\overline{E} \mid E$	$E \mid \overline{E}$
Poisson		22%	10%
Hurdle	0.3	7%	24%
	0.5	8%	20%
	0.7	22%	10%
Zero-inflated	0.3	9%	26%
	0.5	15%	13%
	0.7	29%	7%

Tableau 3 : Taux de mauvaises prédictions de l'événement E par validation croisée

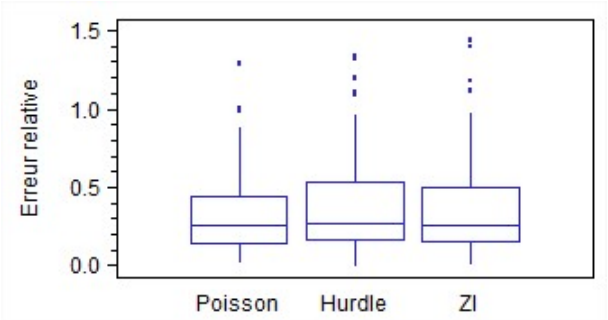

Figure 3 : Distribution des erreurs relatives

Nous constatons que la régression de Poisson ne prédit pas de manière satisfaisante l'événement E lorsqu'il a lieu (erreur de type $\overline{E} \mid E$), mais le nombre d'événements E prédits à tort est relativement faible (erreur de type $E \mid \overline{E}$). Les modèles hurdle et zero-inflated sont plus efficaces pour la prédiction de E, mais en contrepartie, il y aura de nombreuses fausses alertes. Il est clair qu'il est plus grave pour nous de ne pas prédire E alors qu'il se réalise que de prédire E à tort, cela signifierait ne pas alerter le recruteur alors que sa campagne risque d'échouer. Le modèle hurdle semble fournir le meilleur équilibre entre les deux types d'erreurs (avec par exemple un seuil égal à 0.5) étant donné la remarque précédente.

Si E n'est pas prédit, nous estimons le nombre de candidatures reçues. La figure 3 présente les erreurs de prédiction relatives (erreur absolue divisée par le nombre de CV réel) pour chacun des modèles. Bien que les erreurs médianes soient comparables pour les trois modèles, la régression de Poisson apparaît plus robuste avec une distribution légèrement plus resserrée. La régression hurdle, performante pour la prédiction de l'événement E, semble moins efficace pour l'estimation du nombre de CV, à l'inverse de la régression de Poisson. Le modèle zero-inflated semble un intermédiaire entre les deux. Il est probable que nous ayons affaire à un problème de sur-apprentissage étant donné le nombre élevé de paramètres introduits dans ces deux modèles.

5. Conclusions et perspectives

Les régressions de type hurdle et zero-inflated nous apportent une alternative parfaitement adaptée à notre problématique grâce à la modélisation explicite d'un processus en deux étapes. D'une efficacité supérieure à la régression de Poisson standard pour la prédiction des alertes à la première étape, elles se révèlent moins performantes pour la prédiction des effectifs. Nous envisageons de tester une régression de Poisson avec classes latentes (Wedel et al., 1993) car nous suspectons la présence d'une distribution mélangée.
Nous nous sommes basés sur les données d'un seul site d'emploi et avons choisi le seuil d'alerte de manière arbitraire car l'objectif était de mettre en place une méthode de prédiction du nombre de candidatures en deux temps. L'objectif pour la suite est de la mettre en œuvre à plus grande échelle (multi-sites). Afin d'améliorer le pouvoir prédictif du modèle, nous introduirons de nouveaux facteurs explicatifs. Pour une approche orientée sur l'interprétation de ces facteurs, nous envisageons de mettre en place une régression généralisée PLS (Bastien et al., 2005) adaptée au modèle poissonnien afin de pouvoir étudier simultanément toutes les variables même en cas de fortes corrélations.

Bibliographie

[1] Bastien, P., Esposito Vinzi, V. et Tenenhaus, M. (2005) PLS generalised linear regression. *Computational Statistics and Data Analysis*, 48, 17–46.
[2] Consul, P. C. et Famoye, F. (1992) Generalized Poisson regression model. *Communications in Statistics, Theory and Methods*, 21, 89–109.
[3] Famoye, F. et Singh, K. P. (2006) Zero-Inflated Generalized Poisson Regression Model with an Application to Domestic Violence Data, *Journal of Data Science*, 4, 117–130.
[4] Lambert, D. (1992) Zero-Inflated Poisson Regression, With an Application to Defects in Manufacturing. *Technometrics*, 34, 1–14.
[5] Liu, W. et Cela, J. (2008) Count Data Models in SAS. *Proceedings of SAS Global Forum*, paper 371–2008.
[6] McCullagh, P. et Nelder, J. A. (1989) *Generalized Linear Models*, London : Chapman and Hall.
[7] Mullahy, J. (1986) Specification and testing of some modified count data models. *Journal of Econometrics*, 33, 341–365.
[8] Wedel, M., Desarbo, W. S., Bult, J. R. et Ramaswamy, V. (1993) A latent class poisson regression model for heterogeneous count data. *Journal of Applied Econometrics*, 8, 397–411.