Modèles linéaires généralisés sur données de comptages

Master 2 MIND

Ryma Lakehal

Faculté des sciences
Université de Montpellier

8 novembre 2020

Plan

Introduction

Le modèle linéaire généralisé

Données `Parastism`
- Présentation des données `Parastism`
- Visualisation des `Parastism`

Modèle linéaire classique
- Normalité des résidus
- Homoscédasticité

Régression de Poisson
- Ajustement du modèle de régression linéaire de Poisson aux donnéesParastism

Régression binomiale négative
- Ajustement du modèle de régression binomiale négative aux donnéesParastism
- Calculs des effets marginaux

conclusion

- Régression de Poisson
- Régression binomiale négative

Pourquoi ces modèles ?

- Les modèles linéaires classiques ne sont pas adaptés pour analyser des variables à expliquer (ou réponses) de type "comptage".

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$
$$\varepsilon \sim N(0, \sigma)$$

- Les données de type comptage ne sont pas distribuées selon une loi Normale.
- La variance des résidus n'est pas constante mais proportionnelle aux comptages moyens prédits par le modèle.

$$
\begin{aligned}
y_i &\sim N(\mu_i, \sigma) \\
\mathrm{E}(Y|X) &= \mu \\
\mu_i &= \beta_0 + \beta_1 x_i
\end{aligned}
$$

Ces modèles sont constitués de trois éléments :

- un prédicteur linéaire, $\eta = \mathsf{X}\boldsymbol{\beta}$
- une distribution de probabilité de la famille exponentielle $y_i \sim \mathrm{Prob}(\mu_i)$
- une fonction de lien $\eta_i = g(\mu_i)$

Présentation des données `Parastism`

- 196 fruits contient au moins un parasitoïde vivant ou éradiqué
- 63 fruits non infectés "contrôlés"
- 67 fruits avec le parasite *A. melinus* sans sucre
- 66 fruits infectés du même parasite plus le sucre
- au total : 949 parasitoïdesvivants et 365 parasites éliminés contenant un œuf ou larves d'Aphytis.

	Treatment	Fruit	Alive	Parasitized
0	Releases_sugar	1	4	8
1	Releases_sugar	2	0	3
2	Releases_sugar	3	4	3
3	Releases_sugar	4	2	2
4	Releases_sugar	5	1	1

	Fruit	Alive	Parasitized
count	196.000000	196.000000	196.000000
mean	33.188776	4.841837	1.862245
std	18.944185	6.140282	2.955465
min	1.000000	0.000000	0.000000
25%	17.000000	2.000000	0.000000
50%	33.000000	3.000000	1.000000
75%	49.250000	5.000000	2.000000
max	67.000000	37.000000	23.000000

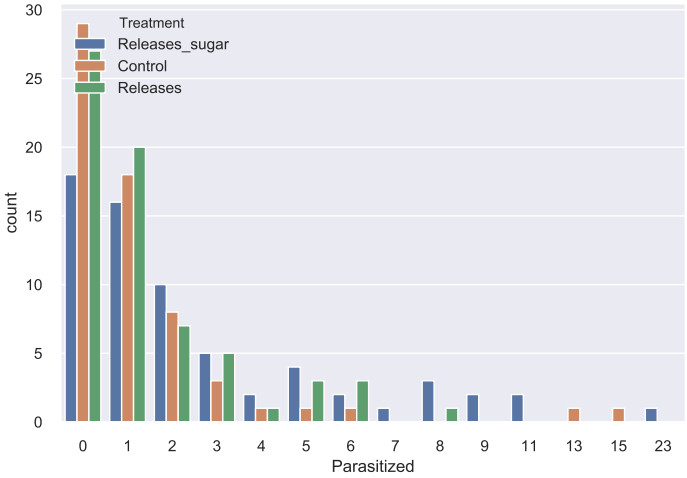

FIGURE – Histogramme des traitements en fonction de la variable Parasitized

Le modèle :

$$Parasitized_i = \beta_0 + \beta_1 Control_i + \beta_2 Realises_i + \beta_3 Realises\_sugar_i + \varepsilon_i$$
$$\varepsilon \sim N(0, \sigma) \tag{3}$$

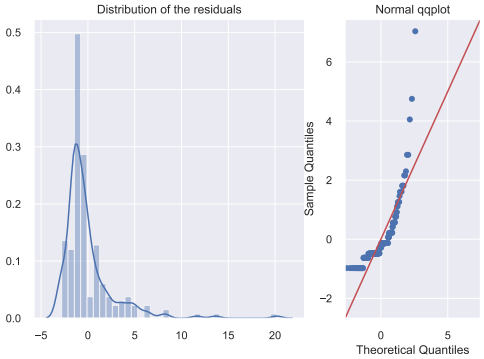

FIGURE – Distribution et normal Q-Q plot des résidus du modèle linéaire ajusté

Homoscédasticité - scale-location plot

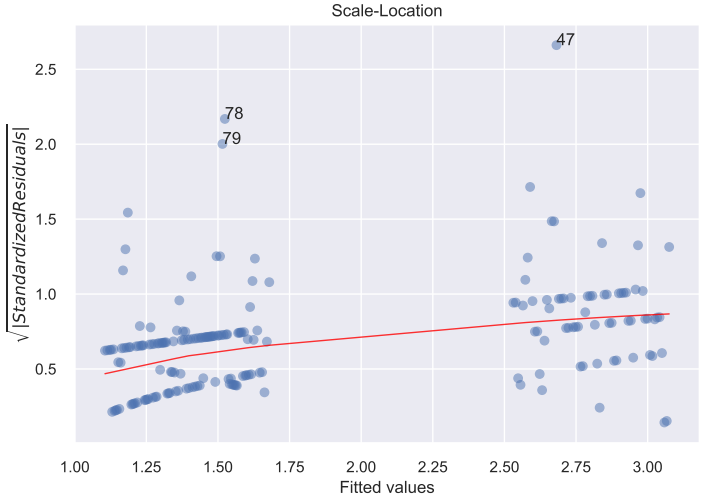

FIGURE – scale-location plot pour vérifier l'homoscédasticité

Le modèle est :

$$
\begin{aligned}
\mathit{Parasitized}_i &\sim \mathit{Poisson}(\mu_i) \\
\mathrm{E}(\mathit{Parasitized} | \mathit{Treatment}) &= \mu \\
\mu_i &= \exp(\eta_i) \\
\eta_i &= \beta_0 + \beta_1 \mathit{Control}_i + \beta_2 \mathit{Realises}_i + \beta_3 \mathit{Realises\_sugar}_i
\end{aligned}
$$

L'ajustement est réalisé à l'aide de la fonction **Poisson.fit** du module **statsmodels**

Régression de Poisson

Generalized Linear Model Regression Results

Dep. Variable:	Parasitized	No. Observations:	196
Model:	GLM	Df Residuals:	193
Model Family:	Poisson	Df Model:	2
Link Function:	log	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-460.45
Date:	Sun, 08 Nov 2020	Deviance:	595.68
Time:	03:02:42	Pearson chi2:	818.
No. Iterations:	5		
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	0.3112	0.108	2.886	0.004	0.100	0.523
C(Treatment)[T.Releases]	0.0274	0.149	0.184	0.854	-0.265	0.320
C(Treatment)[T.Releases_sugar]	0.7195	0.131	5.513	0.000	0.464	0.975

Le ratio residual deviance/ddl est égal à 5304.4/193, soit 27,48. Ce ratio est très largementsupérieur à 1 et permet de mettre en évidence la présence d'une surdispersion

$$Parasitized_i \sim NB(\mu, k)$$
$$\mathrm{E}(Parasitized | Treatment) = \mu$$
$$\mu_i = \exp(\eta_i)$$
$$\eta_i = \beta_0 + \beta_1 Control_i + \beta_2 Realises_i + \beta_3 Realises\_sugar_i$$

Generalized Linear Model Regression Results

Dep. Variable:	Parasitized	No. Observations:	196
Model:	GLM	Df Residuals:	193
Model Family:	NegativeBinomial	Df Model:	2
Link Function:	log	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-355.42
Date:	Sat, 07 Nov 2020	Deviance:	228.53
Time:	22:45:25	Pearson chi2:	290.
No. Iterations:	5		
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	0.3112	0.166	1.877	0.061	-0.014	0.636
C(Treatment)[T.Releases]	0.0274	0.230	0.119	0.905	-0.424	0.479
C(Treatment)[T.Releases_sugar]	0.7195	0.219	3.282	0.001	0.290	1.149

Calculs des effets marginaux

Les effets marginaux sont utilisées pour décrire l'impactvd'un prédicteur sur la variable à expliquer.
On s'intéresse aux effects marginaux à la moyenne, pour ce faire, on a la fonction **.get_margeff()** de la bibliothèque **Statsmodels**.

NegativeBinomial Marginal Effects

Dep. Variable:	Parasitized
Method:	dydx
At:	mean

	dy/dx	std err	z	P>|z|	[0.025	0.975]
C(Treatment)[T.Releases]	0.0481	0.438	0.110	0.913	-0.811	0.907
C(Treatment)[T.Releases_sugar]	1.2632	0.429	2.948	0.003	0.423	2.103

La valeur de **Releases_sugar** est 1.26 ce qui peut être interprété que quand la valeur de **Releases_sugar** augmente d'une unité, la probabilité des parasitoïdes éliminé ou le taux de parasitism augmente de 126%.

Conclusion

En appliquant les différents modèles de modélisation, nous avons constaté que le modèle linéaire généralisé de distribution binomiale négative est celui qui s'adapte le mieux à notre jeu de données **Parastism**. Et de ce dernier modèle, on conclut que les provisions de sucre aident les parasitoïdes à maintenir leurs réserves de sucre ce qui fera augmenter leurs fécondités et ainsi le taux de parasitisme (`Parasitism`)