MATH 60604
Modélisation statistique
§ 4f - Surdispersion

HEC Montréal
Département de sciences de la décision

Extension du modèle de Poisson et surdispersion

- La loi de Poisson n'est pas très flexible car elle possède un seul paramètre qui régule à la fois sa moyenne et sa variance.
- Dans plusieurs cas, cette hypothèse n'est pas valide. Dans l'exemple d'intention d'achat, le rapport déviance sur degrés de liberté était $203{,}27/110 = 1{,}85$, ce qui suggérait que le modèle de Poisson ajusté n'était **pas adéquat** (valeur-p inférieure à 10^{-5}).
- Une des raisons sous-jacentes pour cette piètre performance est la surdispersion, lorsque la variabilité des données de dénombrement est plus élevée que leur moyenne.
- La modèle de régression binomiale négative est souvent utilisé dans ces cas de figure.

Loi binomiale négative

- La loi binomiale négative est une loi de probabilité pour une variable entière dotée de deux paramètres.
- On considère la paramétrisation la plus courante pour la modélisation. La fonction de masse est

$$\mathrm{P}\left(Y = y\right) = \frac{\Gamma(y + 1/k)}{\Gamma(y+1)\Gamma(1/k)} \left(\frac{1/k}{1/k + \mu}\right)^{1/k} \left(\frac{\mu}{1/k + \mu}\right)^{y}$$

  pour $y = 0, 1, 2, 3, \ldots$, où Γ dénote la fonction gamma. Les deux paramètres de la loi sont positifs, $\mu > 0$ et $k > 0$.
- La moyenne théorique et la variance sont

$$\mathrm{E}\left(Y\right) = \mu, \qquad \mathrm{Var}\left(Y\right) = \mu + k\mu^2.$$

- La variance d'une variable binomiale négative est donc supérieure à sa moyenne.

Régression binomiale négative

- Dans la régression binomiale négative, on postule que la variable réponse Y suit une loi binomiale négative avec fonction de liaison logarithmique

$$g\{\mathrm{E}\,(Y_i)\} = \ln\{\mathrm{E}\,(Y_i)\} = \beta_0 + \beta_1 \mathrm{X}_{i1} + \ldots + \beta_p \mathrm{X}_{ip}.$$

- De manière équivalente, cela revient à assumer que la ie observation, Y_i, suit une loi binomiale négative de moyenne

$$\mathrm{E}\,(Y_i) = \exp(\beta_0 + \beta_1 \mathrm{X}_{i1} + \ldots + \beta_p \mathrm{X}_{ip})$$

- Ainsi, l'interprétation des paramètres se fait de manière identique à la régression de Poisson.
- En revanche, la régression binomiale négative a un autre paramètre, k. Ce dernier est supposé identique pour chaque observation. Il ne dépend donc pas des variables explicatives.

Tangente mathématique: À proprement parler, la régression binomiale négative ne fait pas partie des modèles linéaires généralisés parce que cette loi fait partie de la famille de dispersion exponentielle. On peut ajuster le modèle par maximum de vraisemblance et la machinerie développée pour les GLM s'applique néanmoins.

Régression binomiale négative avec `proc genmod`

La seule différence avec le modèle de Poisson est la loi des réponses, viz. `dist=negbin`.

Code SAS code pour ajuster une régression binomiale négative

```
proc genmod data=modstat.intention;
class educ revenu;
model nachat=sexe age revenu educ statut
      fixation emotion / dist=negbin link=log
      lrci type3;
run;
```

Dans R, la paramétrisation de `MASS:glm.nb` est telle que $\theta = 1/k$.

Qualité de l'ajustement pour le modèle binomiale négative

Critères d'évaluation de l'adéquation			
Critère	**DDL**	**Valeur**	**Valeur/DDL**
Ecart	110	118.2310	1.0748
Déviance normalisée	110	118.2310	1.0748
Khi2 de Pearson	110	119.5504	1.0868
Pearson normalisé X2	110	119.5504	1.0868
Log-vraisemblance		14.7494	
Log-vraisemblance complète		-174.6250	
AIC (préférer les petites valeurs)		371.2501	
AICC (préférer les petites valeurs)		373.6945	
BIC (préférer les petites valeurs)		401.9125	

Statistique LR pour Analyse de Type 3			
Source	**DDL**	**Khi-2**	**Pr > khi-2**
sexe	1	3.80	0.0513
age	1	2.23	0.1350
revenu	2	19.68	<.0001
educ	2	2.11	0.3481
statut	1	2.61	0.1061
fixation	1	35.54	<.0001
emotion	1	12.15	0.0005

La déviance est plus près de un. À cause de l'inflation de la variance, seules les variables `revenu`, `fixation` et `emotion` sont significatives.

Estimés des paramètres pour la régression binomiale négative

Analyse des paramètres estimés du maximum de vraisemblance							
Paramètre		**DDL**	**Estimation**	**Erreur type**	**Rapport de vraisemblance Intervalle de confiance à95%**	**Khi-2 de Wald**	**Pr > khi-2**
Intercept		1	-1.1761	0.9729	-3.1103 0.7640	1.46	0.2267
sexe		1	0.5077	0.2550	-0.0029 1.0155	3.96	0.0465
age		1	-0.0415	0.0281	-0.0990 0.0130	2.18	0.1395
revenu	1	1	1.1053	0.3521	0.4124 1.8148	9.86	0.0017
revenu	2	1	-0.1617	0.3535	-0.8660 0.5377	0.21	0.6473
revenu	3	0	0.0000	0.0000	0.0000 0.0000	.	.
educ	1	1	0.3645	0.3441	-0.3263 1.0500	1.12	0.2895
educ	2	1	0.4386	0.3041	-0.1624 1.0494	2.08	0.1492
educ	3	0	0.0000	0.0000	0.0000 0.0000	.	.
statut		1	-0.3873	0.2369	-0.8593 0.0850	2.67	0.1021
fixation		1	0.6316	0.1056	0.4338 0.8581	35.81	<.0001
emotion		1	0.7570	0.2127	0.3401 1.1902	12.66	0.0004
Dispersion		1	0.5840	0.2119	0.2564 1.1193		

L'estimé du paramètre d'échelle $\widehat{k} = 0.584$. À noter que les intervalles de confiance sont basés sur des rapports de vraisemblance et que leurs conclusions peuvent être différentes de celles tirées à l'aide des valeurs-p du test de Wald; préférer les premières, qui sont plus fiables.

Sélection de modèles

- Le quotient déviance sur degrés de libertés est plus près de un, mais cette comparaison est informelle.
- On pourrait utiliser les critères d'information pour choisir le modèle (le plus petit, le mieux); la régression binomiale négative est préférable selon le AIC et le BIC.

modèle	Poisson	binom. nég.
AIC	392.33	371.25
BIC	420.20	301.91

Loi négative binomiale et Poisson

- Quand k tend vers zéro, la loi négative binomiale devient une loi de Poisson.
- On peut comparer les deux modèles à l'aide d'un test du rapport de vraisemblance (modèles emboîtés).
- On teste les hypothèses $\mathscr{H}_0 : k = 0$, $\mathscr{H}_1 : k \neq 0$ à l'aide de la statistique du rapport de vraisemblance
  - attention! la loi nulle est **irrégulière** parce que k doit être positif; quand $n \to \infty$, il y a une probabilité de 0,5 que la déviance soit exactement égale à zéro et 0,5 qu'elle suive une loi χ^2_1 sous $\mathscr{H}_0$.
- La loi nulle asymptotique est

$$2\{\ell_{\text{negbin}}(\widehat{\mu}_{\text{negbin}}, \widehat{k}) - \ell_{\text{pois}}(\widehat{\mu}_{\text{pois}})\} \dot{\sim} \frac{1}{2}\chi^2_1 + \frac{1}{2}\delta_0;$$

En pratique: si on a pas $\widehat{k} = 0$, on calcule la valeur-p comme d'ordinaire à partir de la loi χ^2_1 et on **divise par deux** la valeur-p pour obtenir la **valeur correcte**.

Test du rapport de vraisemblance (irrégulier)

Pour faire les calculs à la main à l'aide des sorties.

- La "**Log-vraisemblance complète**" donne la valeur de la log-vraisemblance du modèle ajusté, −174,6250 pour la régresion binomiale négative et −186,1639 pour la régression de Poisson.
- La différence est 11,5389 et la statistique du rapport de vraisemblance 23,08.

Test du rapport de vraisemblance (irrégulier)

Code SAS pour le test du rapport de vraisemblance (irrégulier)

```
data valp;
valp=(1 - cdf('chisq', 23.08, 1))/2;
run;
proc print data=valp;
run;
```

- La probabilité qu'une loi χ^2_1 soit plus grande que 23,08 est 1.55×10^{-7}.
- Puisque le problème est irrégulier, on divise cette probabilité par deux et notre valeur-p est 7.7×10^{-8}: le modèle de régression binomiale négative est préférable au modèle Poisson.