La régression Bêta

Une alternative intéressante pour modéliser des proportions

Maxime Lacroix

28 septembre 2018

Mise en contexte

Contexte

- Régression sur une variable réponse tenue entre (0,1)
- Par exemple, un taux ou une proportion
- Régression linéaire "classique" à éviter

Première solution : transformation logit

Première transformation possible

$$\widetilde{y} = log(\frac{y}{1-y}) \quad (1)$$

- Avantage :
  - Les données ne sont plus bornées, la régression linéaire est envisageable
- Désavantages :
  - Interprétation différente
  - Fort potentiel d'hétéroscédasticité
  - Les données sont souvent asymétrique $\rightarrow$ problèmes pour les tests d'hypothèses et les intervalles de confiance.

Solution : Régression Bêta

Brève présentation

- Présentée pour la première fois en 2004 par Ferrari et Cribari-Neto
- Intérêt majeur :
  - La densité bêta prend différentes formes dépendemment des paramètres
  - Densité généralement hétéroscédastique
  - Interprétation semblable à la régression logistique

METTRE DES GRAPHIQUES QUI PROUVENT LE POINT 1

Présentation mathématique

Densité d'une loi bêta

$$f_Y(y) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} y^{\alpha-1}(1-y)^{\beta-1} \tag{2}$$

De l'équation 2, Ferrari et Cribari-Neto ont proposé une nouvelle paramétrisation, en posant :

- $\mu = \frac{\alpha}{\alpha+\beta}$
- $\phi = \alpha + \beta$

Nouvelle paramétrisation

Densité sous la nouvelle paramétrisation

$$f_Y(y) = \frac{\Gamma(\mu\phi)}{\Gamma((1-\mu)\phi)} y^{\mu\phi-1}(1-y)^{(1-\mu)(\phi-1)} \tag{3}$$

On peut donc dire que $y \sim B(\mu, \phi)$. L'équation 3 nous donne les propriétés suivantes :

- $E(y) = \mu$
- $Var(y) = \frac{\mu(1-\mu)}{1+\phi}$

On appelle d'ailleurs ϕ le paramètre de dispersion.

Modèle de régression

Définition du modèle

On peut maintenant définir le modèle pratiquement comme un GLM, c'est à dire :

Modèle de régression bêta simple

$$g(\mu_i) = x_i^t \beta = \eta_i \tag{4}$$

La fonction de lien $g()$ peut être choisie comme pour un GLM classique, soit en utilisant le logit, le log-log, le Cauchy. C'est au choix de l'utilisateur. De base, le package `betareg` utilise le lien logit.

La variance de y_i est donnée par la formule suivante. On remarque facilement qu'elle dépend de μ_i, donc il y a hétéroscédasticité.

$$VAR(y_i) = \frac{\mu_i(1-\mu_i)}{1+\phi} \tag{5}$$

Paramètre de dispersion non-constant

Smithson et Verkuilen, en 2006, ont proposé un modèle où le paramètre de dispersion est non-constant. On se retrouve donc avec un autre paramètre à estimer. La régression est donnée par :

Modèle de régression bêta avec dispersion changeante

$$g_1(\mu_i) = x_i^t \beta = \eta_{1i} \tag{6}$$

$$g_2(\phi_i) = z_i^t \gamma = \eta_{2i} \tag{7}$$

Estimation des paramètres

Les paramètres sont estimés en maximisant la vraisemblance. La fonction log-vraisemblance est aisémant calculable, elle est donnée par :

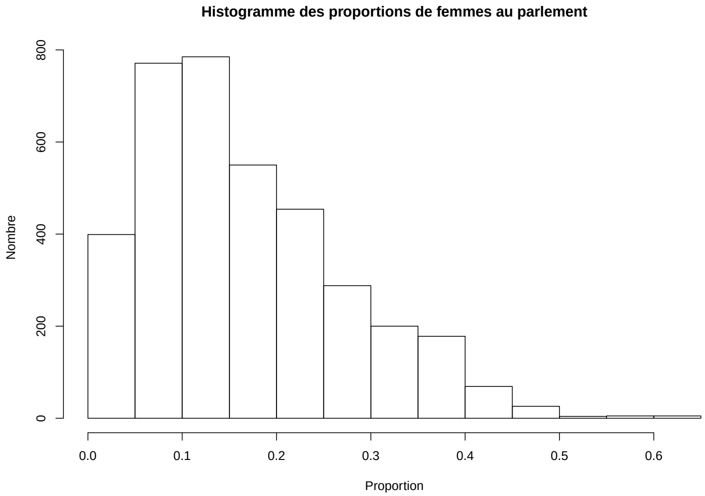
Fonction du log-vraisemblance

$$\begin{aligned} l_i(\mu_i, \phi_i) = & log\Gamma(\phi_i) - log\Gamma(\mu_i\phi_i) - \\ & log\Gamma((1-\mu_i)\phi_i) + (\mu_i\phi_i - 1)log(y_i) + \\ & ((1-\mu_i)\phi_i - 1)log(1-y_i) \end{aligned} \tag{8}$$

Exemple d'utilisation

Jeu de données utilisé

- Titre : Proportion of seats held by women in national parliaments (%)
- Source : https://datahub.io/world-bank/sg.gen.parl.zs#resource-data
- Nombre de données : 3917
- Variables :
  - Nom du pays
  - Code ISO du pays
  - Continent
  - Année
  - Proportion


Histogramme des proportions de femmes au parlement
Nombre
0
200
400
600
800
0.0
0.1
0.2
0.3
0.4
0.5
0.6
Proportion

Implémentation en R

Supposons que l'on veut prédire la proportion de femmes en fonction du continent. Comme il s'agit d'une proportion, on peut utiliser la régression bêta. Le package `betareg` nous permet de faire exactement ça. L'écriture à utiliser est la suivante :

```
betareg(formula, data, subset, na.action, weights, offset,
  link = c("logit", "probit", "cloglog",
           "cauchit", "log", "loglog"),
  link.phi = NULL, type = c("ML", "BC", "BR"),
  control = betareg.control(...), model = TRUE,
  y = TRUE, x = FALSE, ...)
```

Premier modèle : Lien logit et dispersion fixe

```
library(betareg)
mod1 <- betareg(Value~continent,
                link = "logit",
                data=dat_prop)
summary(mod1)
```

Résumé du modèle

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-1.6992958	0.0215978	-78.678997	0
continentAmericas	0.1839680	0.0323398	5.688597	0
continentAsia	-0.1850590	0.0323821	-5.714861	0
continentEurope	0.4418774	0.0296008	14.927878	0
continentOceania	-0.5225864	0.0625121	-8.359764	0

	Estimate	Std. Error	z value	Pr(>|z|)
(phi)	12.86361	0.2964138	43.39748	0

Deuxième modèle : Lien logit et dispersion changeante

```r
mod2 <- betareg(Value~continent|continent,
                link = "logit",
                link.phi = "log",
                data=dat_prop)
summary(mod2)
```

Résumé du modèle

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-1.6752997	0.0238124	-70.354140	0.00e+00
continentAmericas	0.1390195	0.0341478	4.071108	4.68e-05
continentAsia	-0.2175871	0.0357081	-6.093497	0.00e+00
continentEurope	0.4106906	0.0316832	12.962428	0.00e+00
continentOceania	-0.4636063	0.0775094	-5.981290	0.00e+00

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	2.4364670	0.0432605	56.320765	0.0000000
continentAmericas	0.2537546	0.0676070	3.753376	0.0001745
continentAsia	0.1552511	0.0650542	2.386489	0.0170101
continentEurope	0.1774846	0.0631738	2.809463	0.0049624
continentOceania	-0.1322834	0.1194919	-1.107049	0.2682727

Troisième modèle : Lien log et dispersion constante

```
mod3 <- betareg(Value~continent|1,
                link = "log",
                data=dat_prop)
summary(mod3)
```

Résumé du modèle

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-1.8671906	0.0182597	-102.257294	0
continentAmericas	0.1532282	0.0268729	5.701959	0
continentAsia	-0.1586004	0.0278040	-5.704221	0
continentEurope	0.3594905	0.0240928	14.921048	0
continentOceania	-0.4576135	0.0560159	-8.169355	0

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	2.554403	0.0230428	110.8546	0

Comparaison des différents modèles

Les différents modèles nous ont montré différentes choses :

- Peu importe le lien choisi, l'estimation de ϕ est la même si on le choisi constant.
- Comme on maximise la vraisemblance, il est naturel que les coefficients pour μ ne soient pas les mêmes pour les modèles 1 et 2.

Maintenant, on pourrait se demander quel est le meilleur modèle. Plusieurs options s'offrent à nous, que ce soit un test de vraisemblance ou en utilisant le critère de notre choix.

Test du maximum de vraisemblance.