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

Transformation possible

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

- Avantage :
  - Les données ne sont plus bornées, la régression linéaire est envisageable
- Désavantages :
  - Interprétation différente
  - Fort potentiel d'hétéroscédasticité
  - Les données sont souvent asymétriques $\rightarrow$ 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

Présentation mathématique

Densité d'une loi bêta

$$f_Y(y) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} y^{\alpha-1}(1-y)^{\beta-1} \quad y \in (0,1) \quad \alpha, \beta > 0 \tag{2}$$

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 d'une observation 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 :

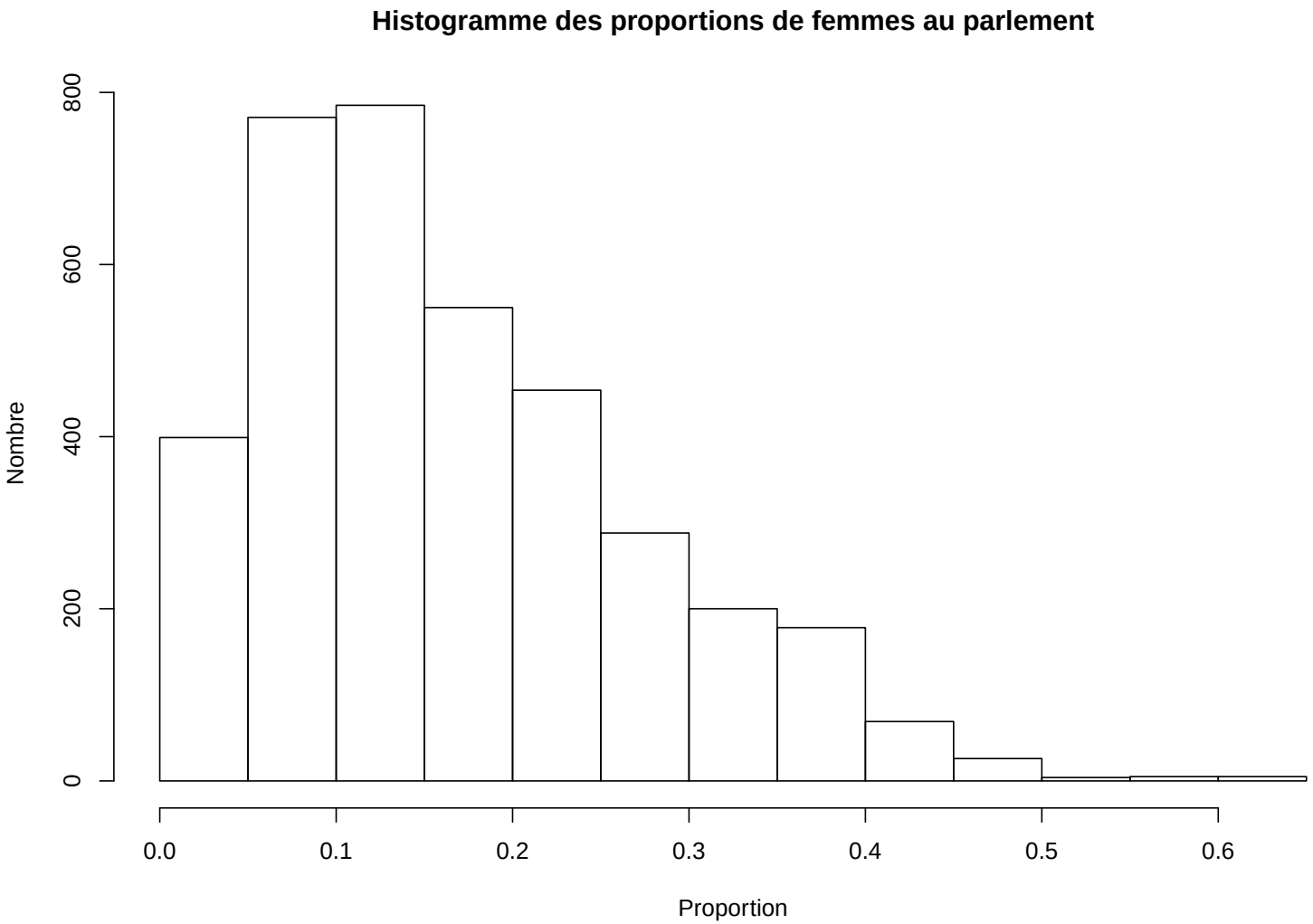
Fonction du log-vraisemblance

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

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
  - Décennie*
  - 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 d'effectuer cette analyse. 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

```
##
## Call:
## betareg(formula = Value ~ continent, data = dat_prop, link = ("logit"))
##
## Standardized weighted residuals 2:
##     Min      1Q  Median      3Q     Max
## -3.8675 -0.5996  0.0200  0.7634  2.8858
##
## Coefficients (mean model with logit link):
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)       -1.69930    0.02160 -78.679  < 2e-16 ***
## continentAmericas  0.18397    0.03234   5.689 1.28e-08 ***
## continentAsia     -0.18506    0.03238  -5.715 1.10e-08 ***
## continentEurope    0.44188    0.02960  14.928  < 2e-16 ***
## continentOceania  -0.52259    0.06251  -8.360  < 2e-16 ***
##
## Phi coefficients (precision model with identity link):
##       Estimate Std. Error z value Pr(>|z|)
## (phi)  12.8636     0.2964    43.4   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Type of estimator: ML (maximum likelihood)
## Log-likelihood:  3750 on 6 Df
## Pseudo R-squared: 0.1299
## Number of iterations: 14 (BFGS) + 2 (Fisher scoring)
```

Deuxième modèle : Lien logit et dispersion changeante

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

Résumé du modèle

```
##
## Call:
## betareg(formula = Value ~ continent | continent, data = dat_prop,
##     link = "logit", link.phi = "log")
##
## Standardized weighted residuals 2:
##     Min      1Q  Median      3Q     Max
## -3.9433 -0.5990  0.0204  0.7282  2.7119
##
## Coefficients (mean model with logit link):
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)       -1.67530    0.02381 -70.354  < 2e-16 ***
## continentAmericas  0.13902    0.03415   4.071 4.68e-05 ***
## continentAsia     -0.21759    0.03571  -6.093 1.10e-09 ***
## continentEurope    0.41069    0.03168  12.962  < 2e-16 ***
## continentOceania  -0.46361    0.07751  -5.981 2.21e-09 ***
##
## Phi coefficients (precision model with log link):
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)        2.43647    0.04326  56.321  < 2e-16 ***
## continentAmericas  0.25375    0.06761   3.753 0.000174 ***
## continentAsia      0.15525    0.06505   2.386 0.017010 *
## continentEurope    0.17748    0.06317   2.809 0.004962 **
## continentOceania  -0.13228    0.11949  -1.107 0.268273
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Type of estimator: ML (maximum likelihood)
## Log-likelihood:  3761 on 10 Df
## Pseudo R-squared: 0.1287
## Number of iterations: 19 (BFGS) + 2 (Fisher scoring)
```

Troisième modèle : Lien log et dispersion constante

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

Résumé du modèle

```
##
## Call:
## betareg(formula = Value ~ continent | 1, data = dat_prop, link = "log")
##
## Standardized weighted residuals 2:
##     Min      1Q  Median      3Q     Max
## -3.8675 -0.5996  0.0200  0.7634  2.8858
##
## Coefficients (mean model with log link):
##                   Estimate Std. Error  z value Pr(>|z|)
## (Intercept)       -1.86719    0.01826 -102.257  < 2e-16 ***
## continentAmericas  0.15323    0.02687    5.702 1.18e-08 ***
## continentAsia     -0.15860    0.02780   -5.704 1.17e-08 ***
## continentEurope    0.35949    0.02409   14.921  < 2e-16 ***
## continentOceania  -0.45761    0.05602   -8.169 3.10e-16 ***
##
## Phi coefficients (precision model with log link):
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  2.55440    0.02304   110.9   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Type of estimator: ML (maximum likelihood)
## Log-likelihood:  3750 on 6 Df
## Pseudo R-squared: 0.131
## Number of iterations: 13 (BFGS) + 2 (Fisher scoring)
```

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 choisi que ϕ est 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.

Quel est le meilleur modèle?

Test du maximum de vraisemblance

On pourrait s'intéresser à connaître l'impact d'ajouter les coefficients de dispersion. En posant :

- H_0 = Modèle simple
- H_1 = Modèle complexe

On peut utiliser le code suivant :

```
library(lmtest)
lrtest(mod1,mod2)
```

Résultat du test

```
## Likelihood ratio test
##
## Model 1: Value ~ continent
## Model 2: Value ~ continent | continent
##   #Df LogLik Df  Chisq Pr(>Chisq)
## 1   6 3750.1
## 2  10 3761.1  4 22.014  0.0001992 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '
```

Choix selon un critère

Tout comme dans les GLM, il est possible de choisir le meilleur en utilisant le critère de notre choix, comme l'AIC par exemple, ou le BIC. Il se peut que le choix du type de lien $g()$ améliore grandement le modèle, dépendemment des données. Le code pour effectuer ce calcul est le suivant :

```
AIC(mod1,mod2,mod3)
```

```
##       df       AIC
## mod1   6 -7488.227
## mod2  10 -7502.241
## mod3   6 -7488.227
```

Méthodes avancées

Contexte

En 2010, Grün, Kosmidis et Zeilis ont publié l'article *Extended Beta Regression in R: Shaken, Stirred, Mixed and Partitioned*, dans lequel ils expliquent différentes méthodes plus complexes pouvant améliorer le modèle. Trois gros thèmes y ressortent, soient :

- Correction de biais dans l'estimation des paramètres
- Les arbres de régression bêta
- Mixtures ou mélanges de régressions bêta

Correction de biais

Dans l'article, on discute le fait que la méthode du maximum de vraisemblance pour estimer les paramètres de ϕ a tendance à sous-estimer l'écart-type des paramètres, ce qui amène des problèmes notamment au niveau de l'inférence que l'on peut faire avec le modèle.

Solution :

- Correction du biais (BC)
- Réduction du biais (BR)

Implémentation en R

Rappel : Écriture de la fonction betareg :

```
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, ...)
```

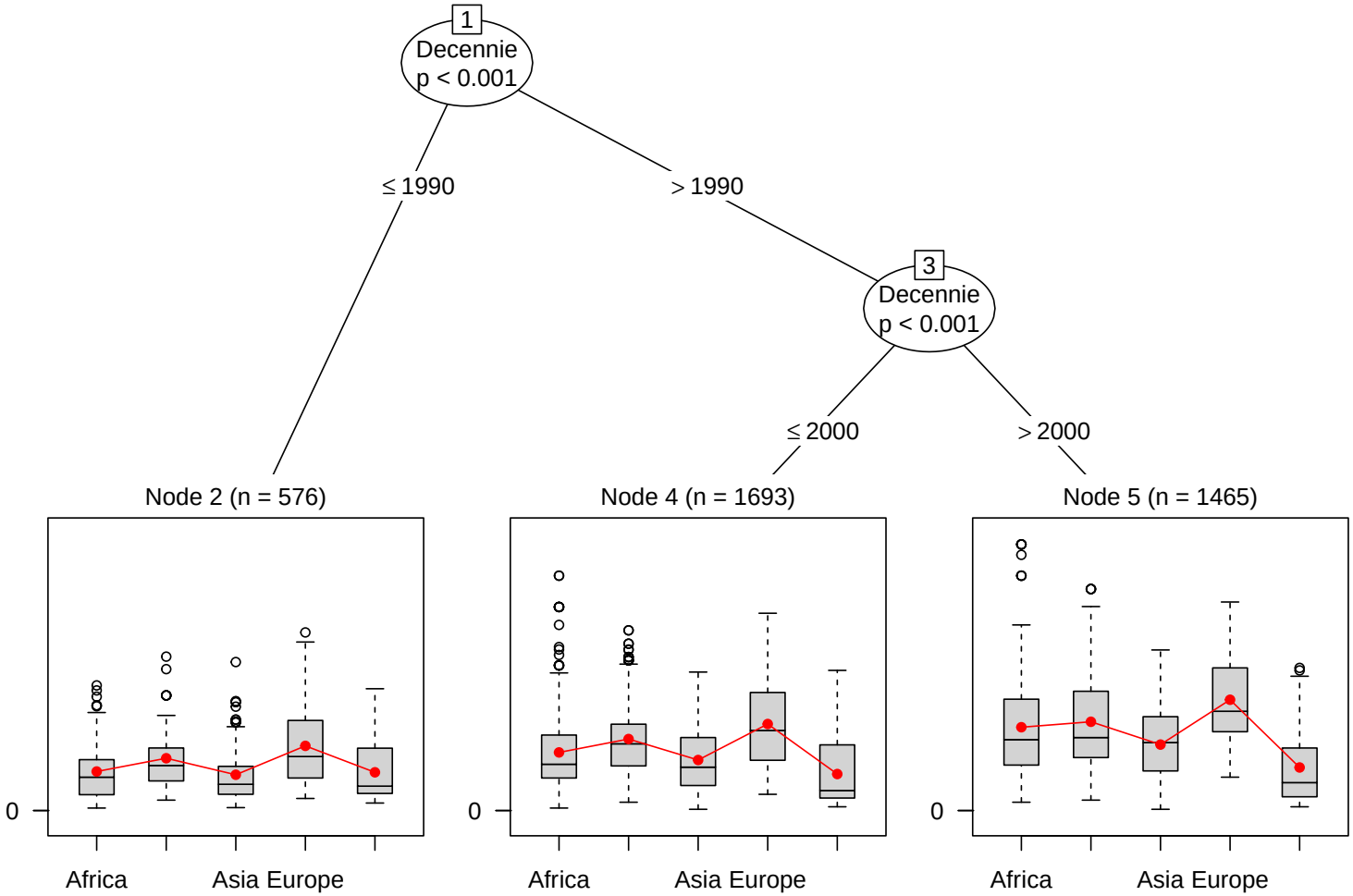
Remarquons le type correspond à la méthode d'estimation choisie.

Arbres de régression bêta

L'idée derrière cette méthode est de se questionner à savoir s'il n'y aurait pas une autre variable qui permettrait de partitionner le modèle. Après avoir partitionné le modèle, on peut estimer les paramètres dans chacune des partitions. Voici un example simple :

```
mod4 <- betatree(Value~continent,
                 ~Decennie,
                 link = "logit",
                 data=dat_prop,
                 minsize=100)
plot(mod4)
```

Résultat



On voit que de faire un modèle différent par décennie semble une option

Mélange de régressions bêta

On pourrait supposer qu'il y a des différences dans différents sous-groupes de l'échantillon, mais qu'il n'existe pas de variables discriminantes. La solution : utiliser `betamix`. Voici un example. Ça revient à créer 3 *clusters* de données et d'estimer un modèle pour chacun des *clusters*.

```
rs_mix <- betamix(accuracy ~ iq,
                  data = ReadingSkills, k = 3,
                  nstart = 10,
                  extra_components = extraComponent(
                  type = "uniform",
                  oef = 0.99, delta = 0.01))
```

Conclusion

Conclusion

- Modèle relativement simple à utiliser
- Possibilité de complexifier les choses rapidement
- Règle beaucoup de problèmes avec la dispersion des données que la régression linéaire n'est pas en mesure de faire.