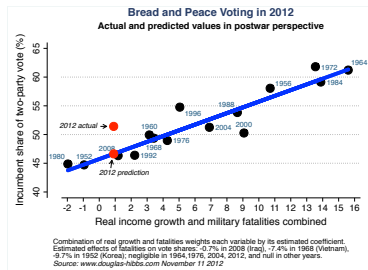


MODÉLISATION GÉNÉRALISÉE

- 1 Modèles linéaires et modèles généralisés
- 2 Estimer des modèles linéaires généralisés dans R
- 3 La modélisation multiniveau, une généralisation du modèle linéaire



Modèles linéaires et modèles généralisés

La régression classique repose sur l'hypothèse de linéarité. Or, les processus étudiés peuvent obéir à bien d'autres formes fonctionnelles. Pour les modéliser, on recourt – dans certains cas – à des modèles linéaires **généralisés**, qui reposent sur :

- des **transformations** du vecteur des prédicteurs
- une **fonction lien**
- une **distribution** de la variable dépendante.

Un cas particulier : la régression logistique

La régression linéaire nécessite que la variable dépendante soit **continue**.

Lorsque la variable dépendante est **binaire**, on recourt à la régression logistique. Il s'agit d'une régression linéaire, dans laquelle on modélise non pas Y mais la probabilité que Y soit égale à 1. Pour faciliter l'estimation, on estime en réalité le **logit** de cette probabilité :

$$p = Pr(Y = 1)$$

$$\log\left(\frac{p}{1-p}\right) = \alpha + \beta + \epsilon$$

Un cas particulier : la régression logistique

Les coefficients des régressions logistiques sont **non linéaires** et sont de ce fait difficiles à interpréter.

On regarde en général la pente autour de la moyenne de la variable indépendante, qu'on peut approximer en **divisant par 4** le coefficient.

Une autre manière d'interpréter le coefficient est de le transformer en **odd ratio**.

Quelques classes de modèles linéaires généralisés

- Le modèle de **Poisson** : pour les données de décompte (et notamment les événements rares). Utilise la transformation logarithmique et une distribution de Poisson.
- Le modèle **binomial-logistique** porte également sur des données de décompte binaire (succès/échec). La transformation est une transformation de type logit, et la distribution est binomiale.
- Le modèle **probit** est un modèle comparable au modèle logit ; on substitue simplement une distribution normale à une distribution logistique. Les résultats sont très comparables.

Quelques classes de modèles linéaires généralisés

- Les modèles **multinomiaux**, ordonnés ou non ordonnés, logit ou probit, généralisent la régression logistique à une variable dépendante polytomique. Ils reposent sur la distribution multinomiale.
- Les modèles **robustes** sont des modèles de régression (linéaires ou logistiques) qui donnent aux erreurs une distribution permettant des valeurs extrêmes – habituellement une distribution t . Les coefficients sont ainsi moins sensibles aux *outliers*.

Estimer les modèles linéaires généralisés

Les modèles généralisés ne sont pas estimés au moyen des MCO mais de la méthode du **maximum de vraisemblance** (maximum likelihood). Il s'agit d'une méthode itérative, qui peut donc être gourmande en ressources.

Plusieurs indicateurs de **goodness of fit** sont employés : l'AIC, le BIC, le log de la vraisemblance (log-likelihood), la déviance. Tous dénotent un meilleur modèle lorsqu'ils s'approchent de zéro.

Estimer des modèles linéaires généralisés dans R

Généralement, `glm` permet d'estimer des modèles linéaires généralisés. Il faut ensuite spécifier le type de modèle et la nature du lien.

```
glm(y ~ x, family = binomial(link = "logit"), data = data)
```

La famille peut être notamment :

- `binomial`
- `gaussian`
- `poisson`

Voir ?family.

Un exemple de régression logistique

Utiliser le script regBES.R. On veut étudier la probabilité de voter travailliste (plutôt que conservateur) au Royaume-Uni en 2010.

```
regBES <- glm(vote ~ gender + age + income + ethnicity, data=BES,  
              family = binomial(link="logit"))
```

```

library(texreg)
screenreg(regBES, single.row = TRUE)

##
## =====
##                               Model 1
## -----
## (Intercept)                2.85 (0.53) ***
## genderFemale                0.07 (0.12)
## age                        -0.02 (0.00) ***
## income5001-10000            -0.82 (0.47)
## income10001-15000          -0.93 (0.46) *
## income15001-20000          -1.18 (0.46) *
## income20001-25000          -1.25 (0.47) **
## income25001-30000          -1.72 (0.48) ***
## income30001-35000          -2.18 (0.49) ***
## income35001-40000          -1.36 (0.49) **
## income40001-45000          -1.83 (0.51) ***
## income45001-50000          -1.58 (0.50) **
## income50001-60000          -1.91 (0.50) ***
## income60001-70000          -1.91 (0.52) ***
## income70001-80000          -1.35 (0.55) *
## income80001-90000          -2.07 (0.62) ***
## income90001 & over         -2.44 (0.51) ***
## ethnicityWhite British     -0.58 (0.22) **
## -----
## AIC                        1619.60
## BIC                        1711.31
## Log Likelihood             -791.80
## Deviance                   1583.60
## Num. obs.                  1206
## =====
## *** p < 0.001, ** p < 0.01, * p < 0.05

```

Utiliser la modélisation multiniveau ?

Utiliser la modélisation multiniveau ?

- Modélisation

Utiliser la modélisation mult niveau ?

- Modélisation
- Multiscale

Utiliser la modélisation multiniveau ?

- Modélisation
- Multiscale
- Variabilité spatiale des relations

Utiliser la modélisation multiniveau ?

- Modélisation
- Multiscale
- Variabilité spatiale des relations
- Contextes

La modélisation multiniveau

La modélisation multiniveau

- Multiniveau

La modélisation multiniveau

- Multiniveau
- Hiérarchique

La modélisation multiniveau

- Multiniveau
- Hiérarchique
- Mixtes

La modélisation multiniveau

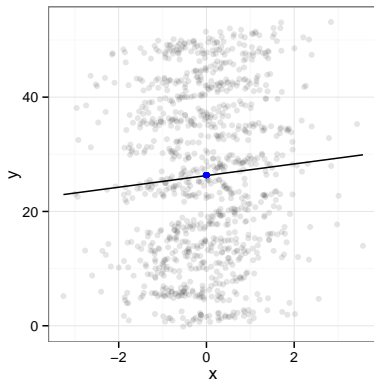
- Multiniveau
- Hiérarchique
- Mixtes
- À coefficients aléatoires

Une généralisation du modèle linéaire

$$y_i = \alpha + \beta x_i + \epsilon_i$$

avec

$$\epsilon_i \sim N(0, \sigma_\epsilon^2)$$



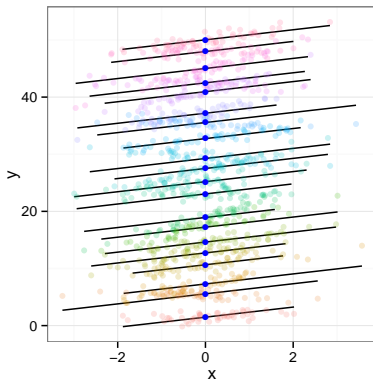
Une généralisation du modèle linéaire

$$y_i = \alpha_{j[i]} + \beta x_i + \epsilon_i$$

avec

$$\alpha_{j[i]} \sim N(\mu, \sigma_\alpha^2)$$

$$\epsilon_i \sim N(0, \sigma_\epsilon^2)$$



Une généralisation du modèle linéaire

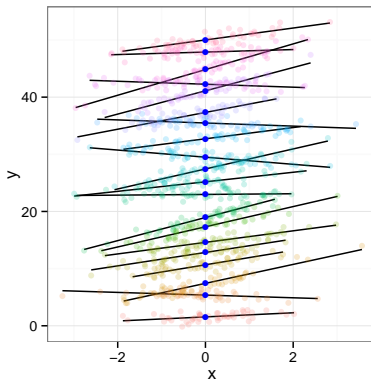
$$y_i = \alpha_{j[i]} + \beta_{j[i]}x_i + \epsilon_i$$

avec

$$\alpha_{j[i]} \sim N(\mu_\alpha, \sigma_\alpha^2)$$

$$\beta_{j[i]} \sim N(\mu_\beta, \sigma_\beta^2)$$

$$\epsilon_i \sim N(0, \sigma_\epsilon^2)$$



Une généralisation du modèle linéaire

Modèle linéaire : hypothèse de constance des coefficients dans toutes les unités

Une généralisation du modèle linéaire

Modèle multiniveau : hypothèse sur la distribution (usuellement normale) de la distribution des coefficients

Une généralisation du modèle linéaire

$$y_i = \alpha_{j[i]} + \beta_{j[i]}x_i + \epsilon_i$$

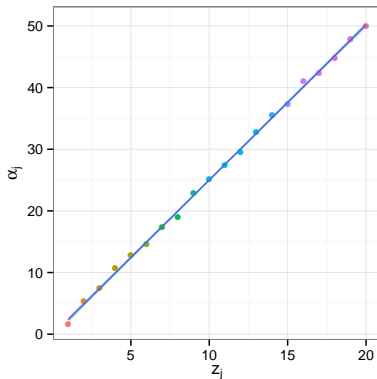
avec

$$\alpha_{j[i]} = \mu_\alpha + \theta z_j + \gamma_j$$

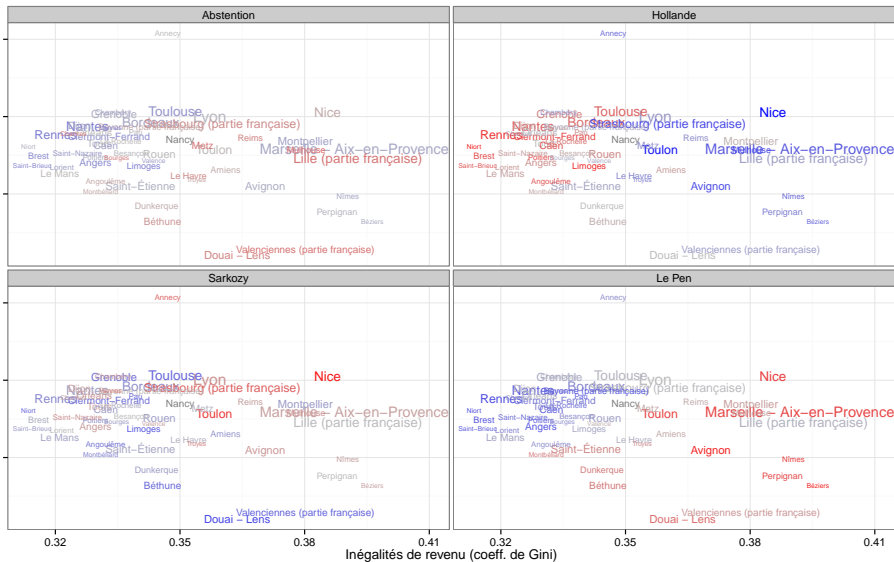
$$\beta_{j[i]} \sim N(\mu_\beta, \sigma_\beta^2)$$

$$\epsilon_i \sim N(0, \sigma_\epsilon^2)$$

$$\gamma_j \sim N(0, \sigma_\gamma^2)$$



Revenu moyen par UC



Résidus standardisés



Utiliser la commande lmer

Le package lme4 offre la commande lmer qui permet, sur le modèle de lm, d'estimer des modèles multiniveaux.

```
lmer(y ~ x + (1 + x | groupe), data = data)
```

Utiliser la commande lmer

Voir le script `regMarseille.R` pour une utilisation de `lmer` : les structures du vote FN à Marseille varient-elles d'un arrondissement à l'autre ?