

# **Validation de modèles de régression en R**

17<sup>e</sup> Colloque du CEF

---

Daniel Schoenig Mégane Déziel

2024-05-01

**Exemple 1: Croissance de l'Épinette de Norvège dans les  
Alpes**

# Exemple 1: Croissance de l'Épinette de Norvège dans les Alpes



## Données

`gutten.rds`

## Références

- Guttenberg, A. R. von. (1915). Wachstum und Ertrag der Fichte im Hochgebirge. Franz Deuticke. <https://doi.org/10.5962/bhl.title.15664>
- Zeide, B. (1993). Analysis of Growth Equations. *Forest Science*, 39(3), 594–616. <https://doi.org/10.1093/forestscience/39.3.594>
- Robinson, A. P., & Hamann, J. D. (2011). *Forest analytics with R: An introduction*. Springer.

Image: Wikimedia (Michela Modena)

# Exemple 1: Croissance de l'Épinette de Norvège dans les Alpes



Image: Wikimedia (Michela Modena)

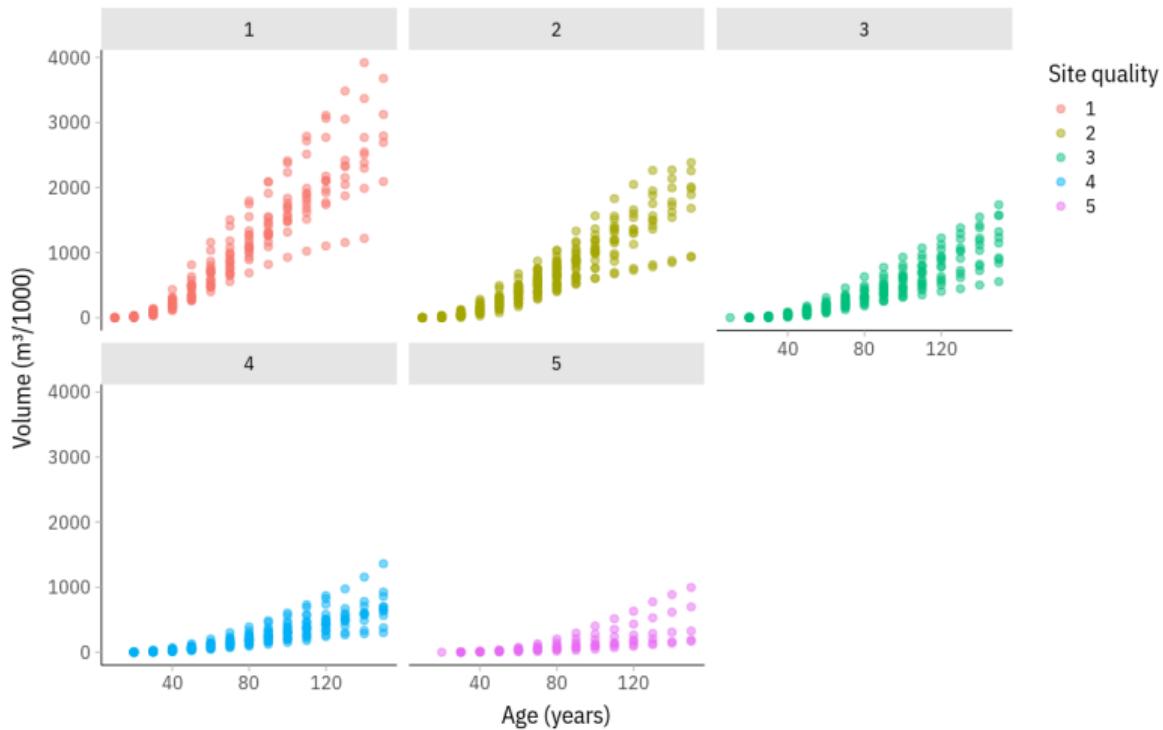
## Données

`gutten.rds`

## Variables

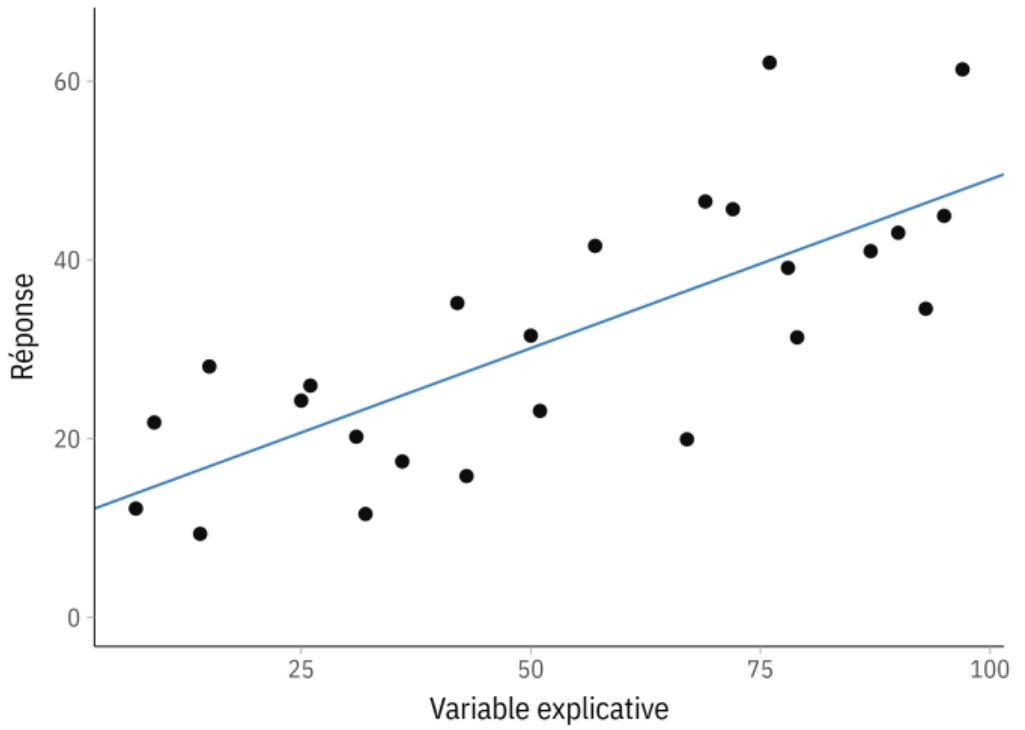
- `quality`: indice de qualité de site, de 1 (le meilleur) à 5 (le pire) ;
- `site`: identité du site ;
- `tree`: identité de l'arbre dans le site ;
- `age.base` : âge de l'arbre déterminé au niveau du sol (années) ;
- `height` : hauteur de l'arbre (m) ;
- `dbh.cm` : diamètre de l'arbre à hauteur de poitrine (cm) ;
- `age.bh` : âge de l'arbre à hauteur de poitrine (années) ;
- `volume` : volume de l'arbre ( $10^{-3} \text{ m}^3$ ) ;
- `tree.id`: identité unique de l'arbre.

# Exemple 1: Croissance de l'Épinette de Norvège dans les Alpes

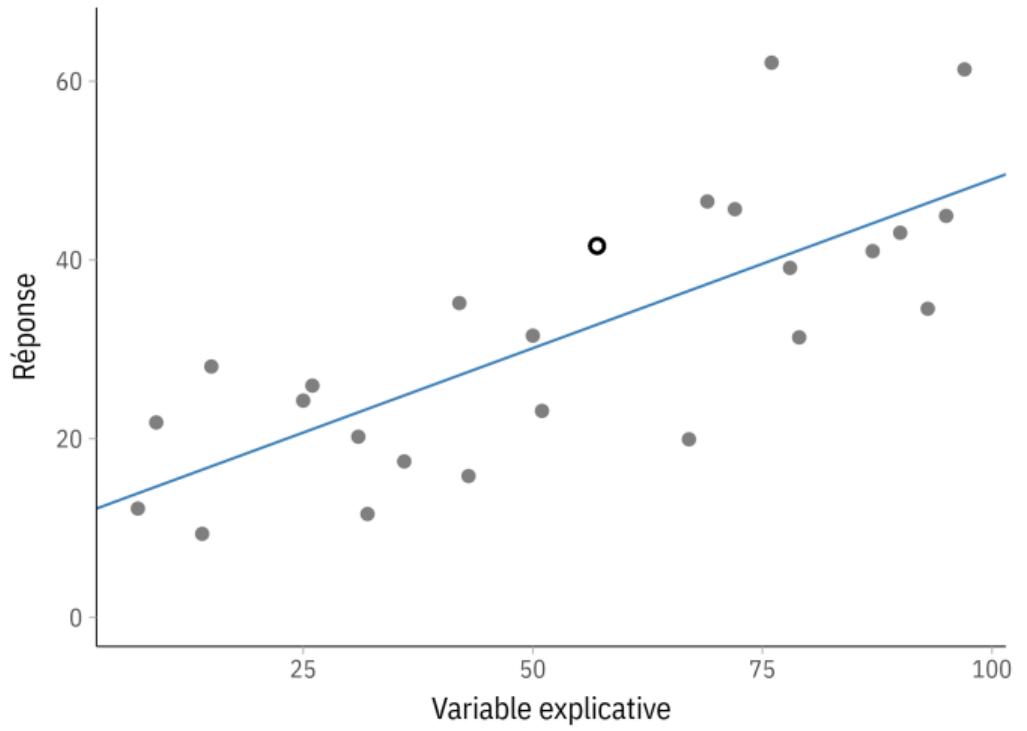


## Résidus quantiles

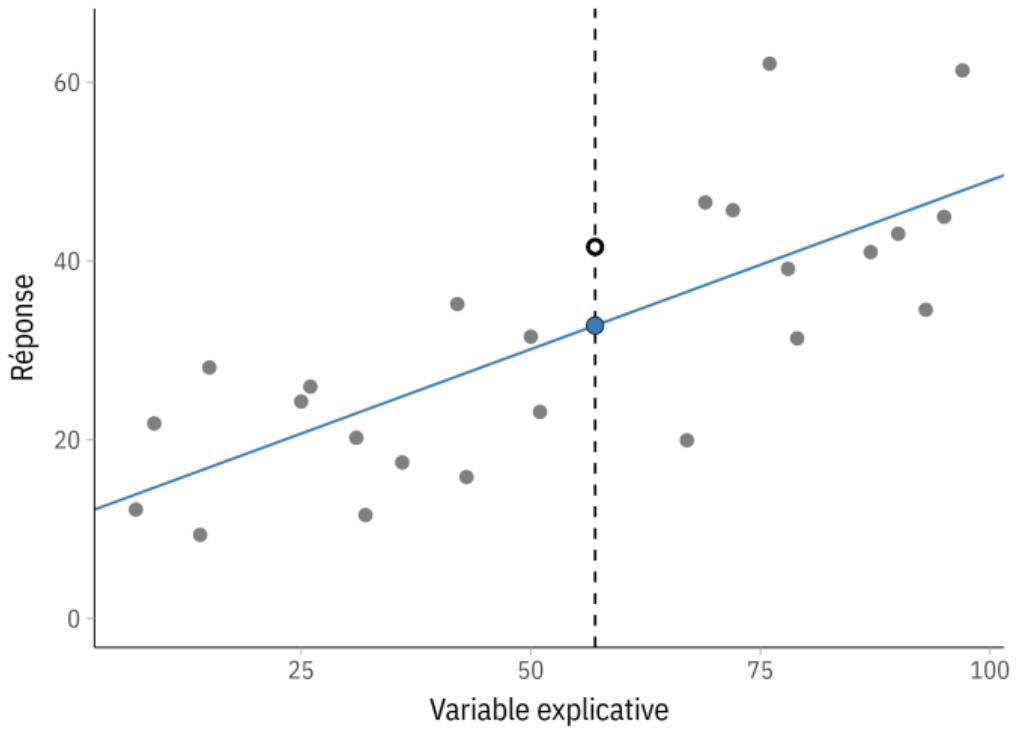
# Résidus



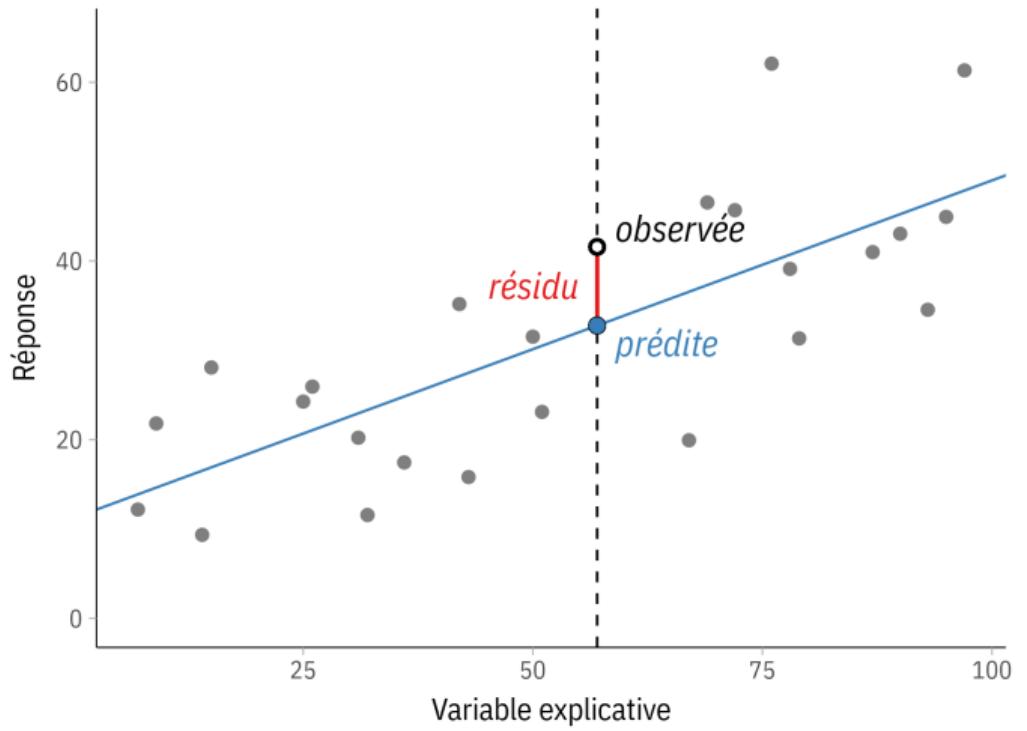
# Résidus



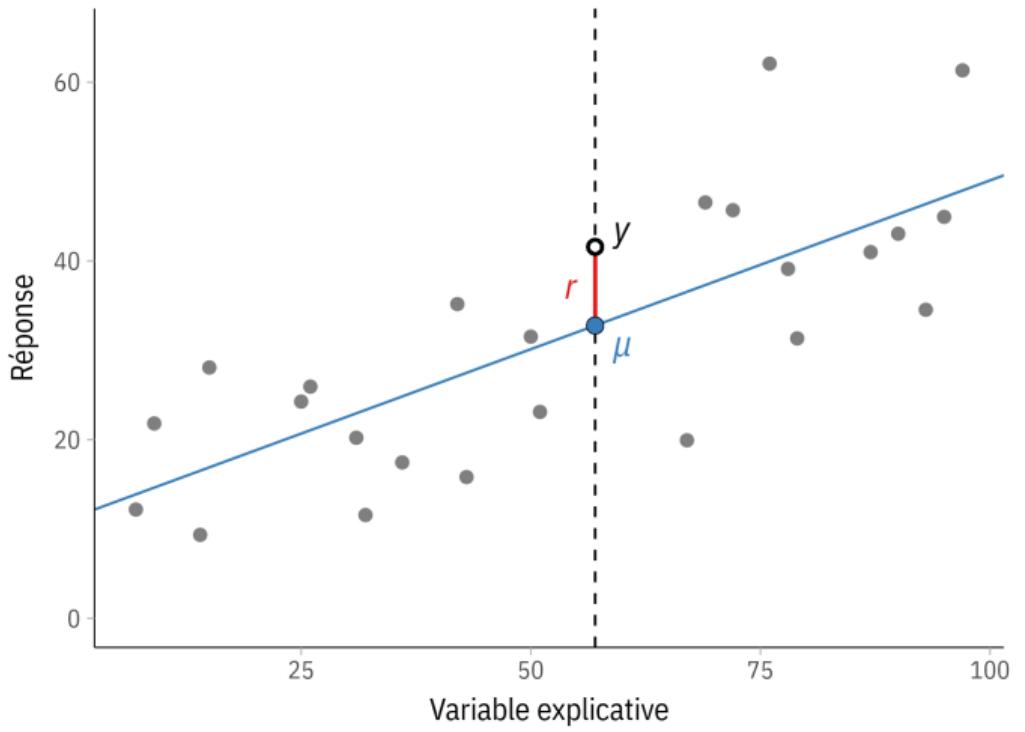
# Résidus



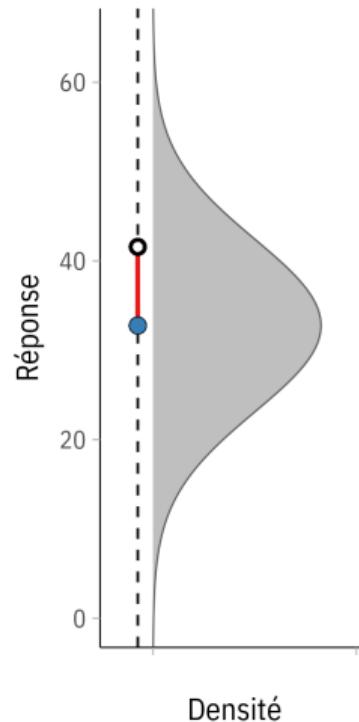
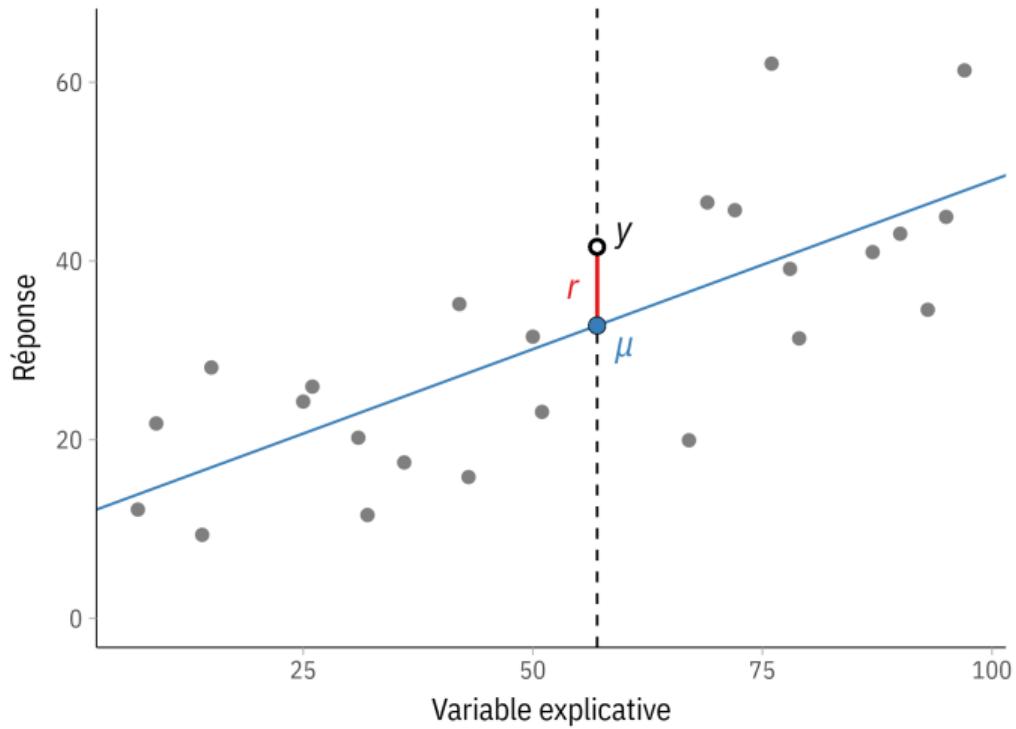
# Résidus



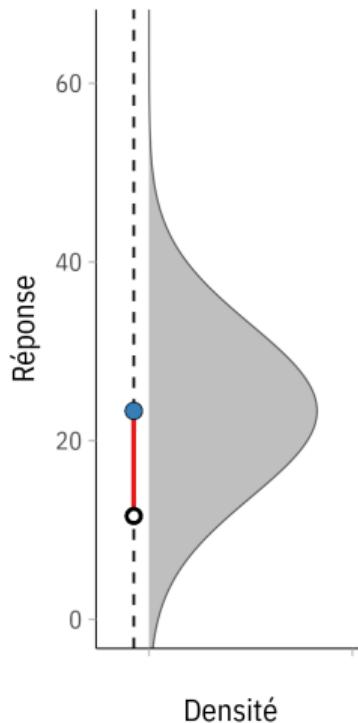
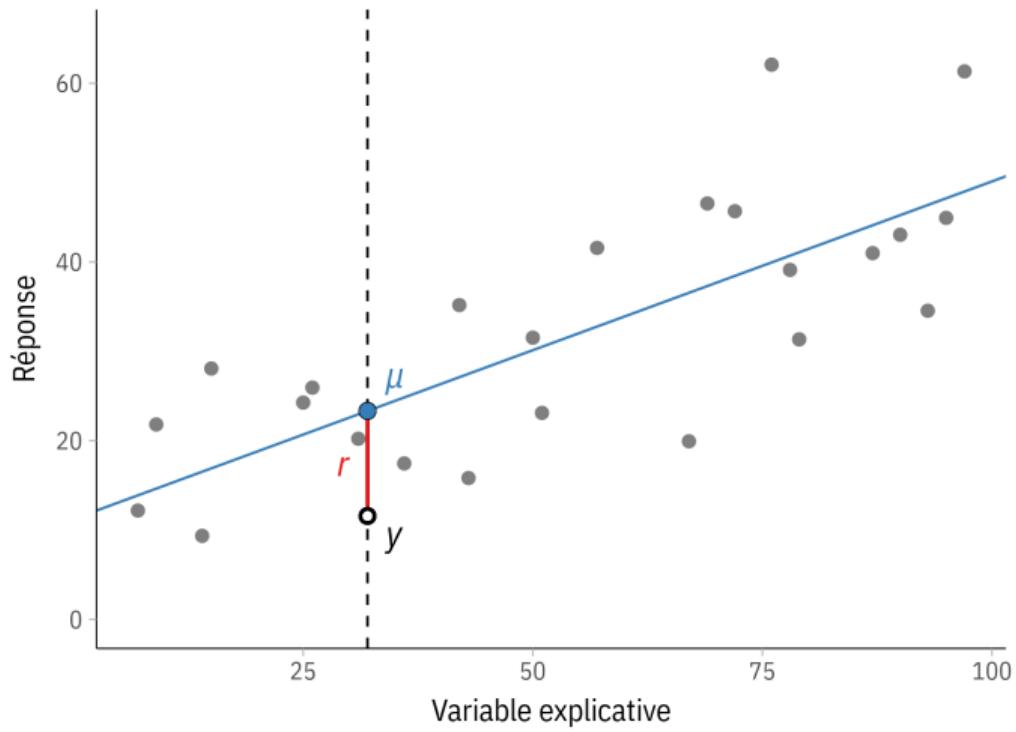
# Résidus



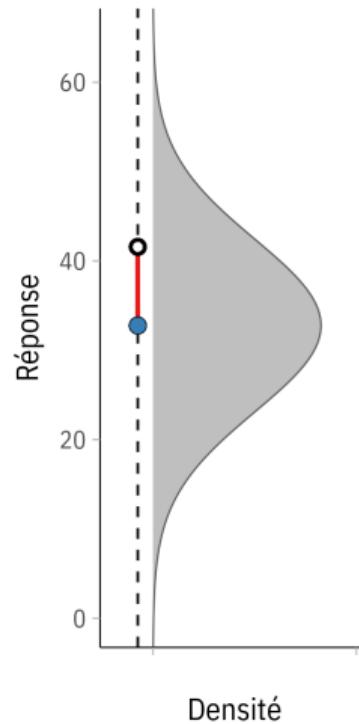
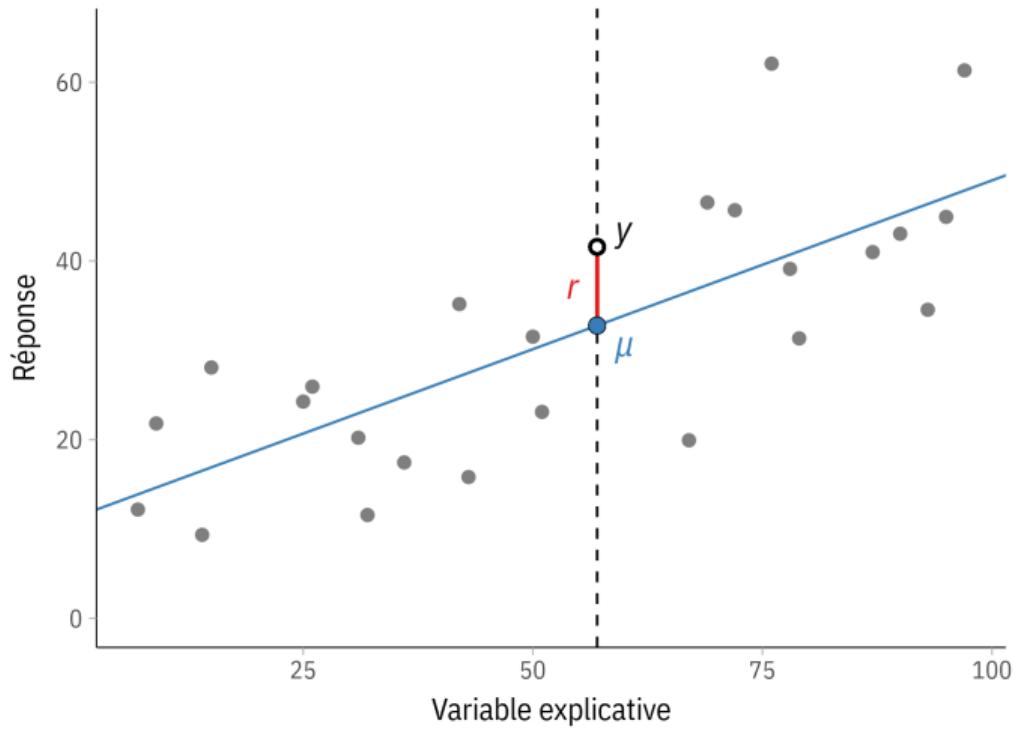
# Résidus



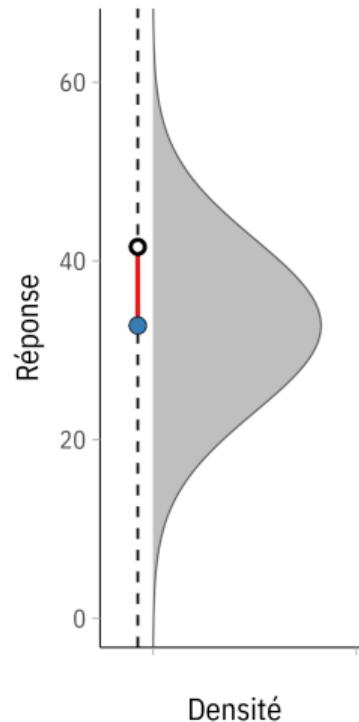
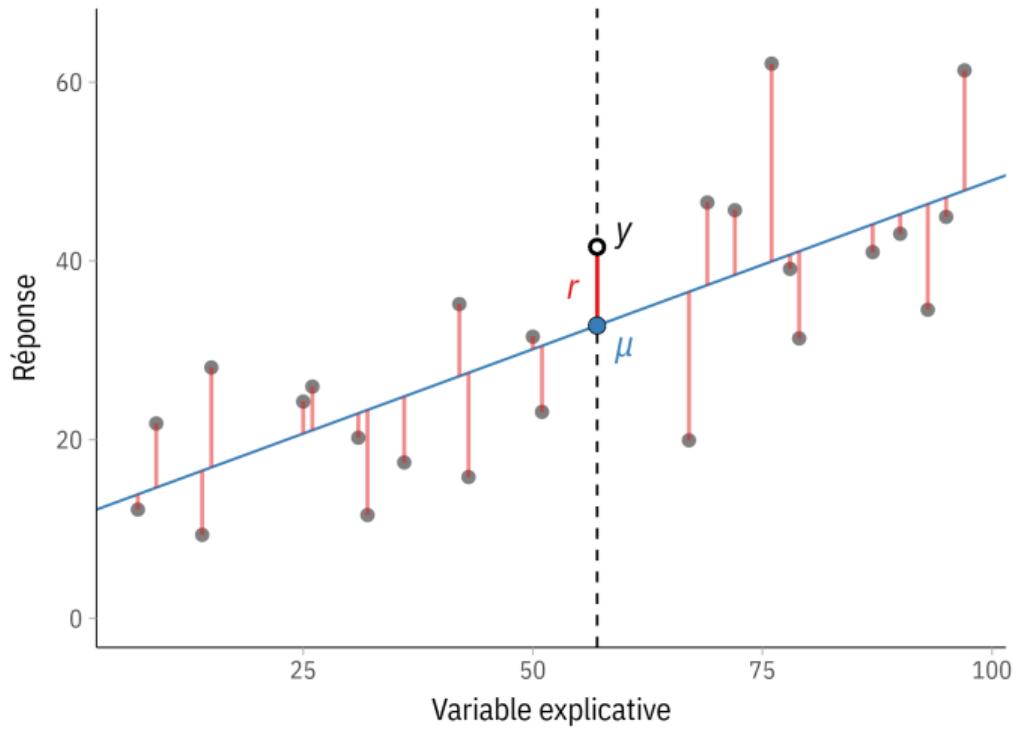
# Résidus



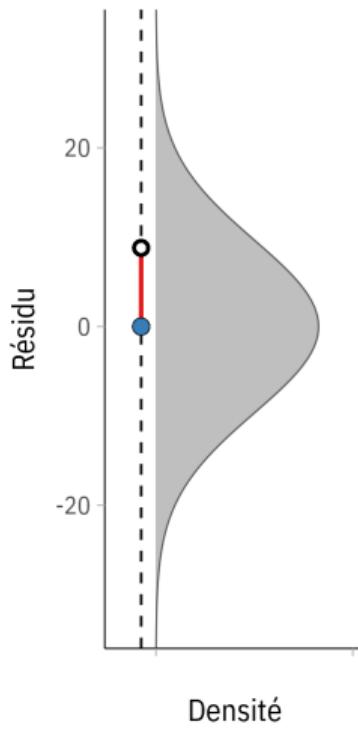
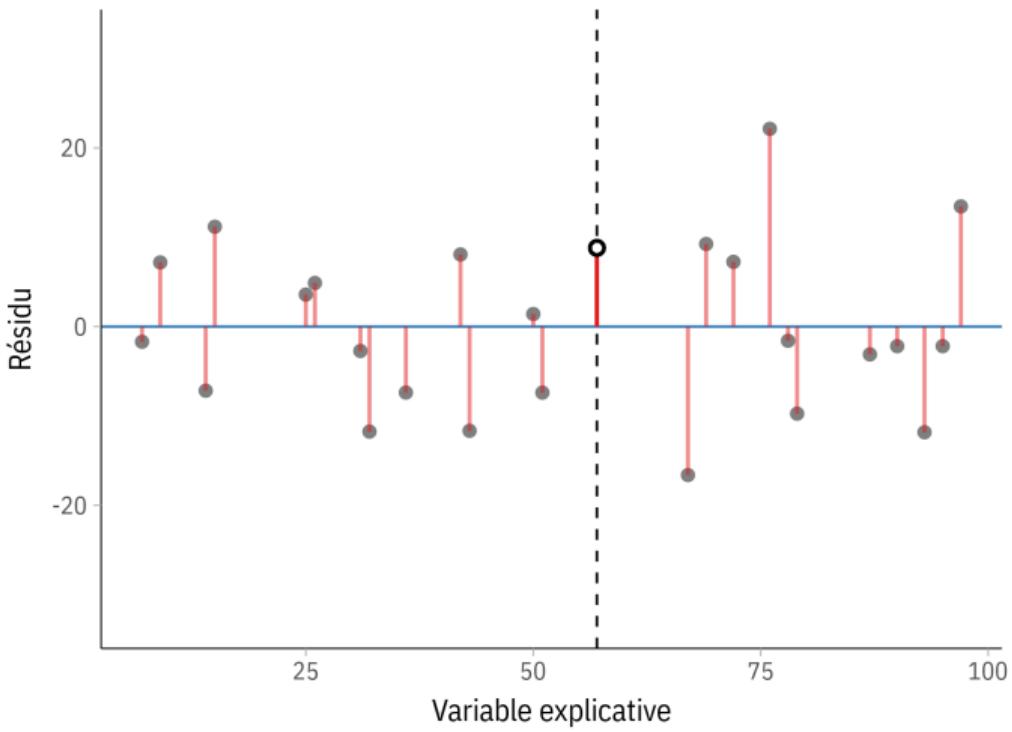
# Résidus



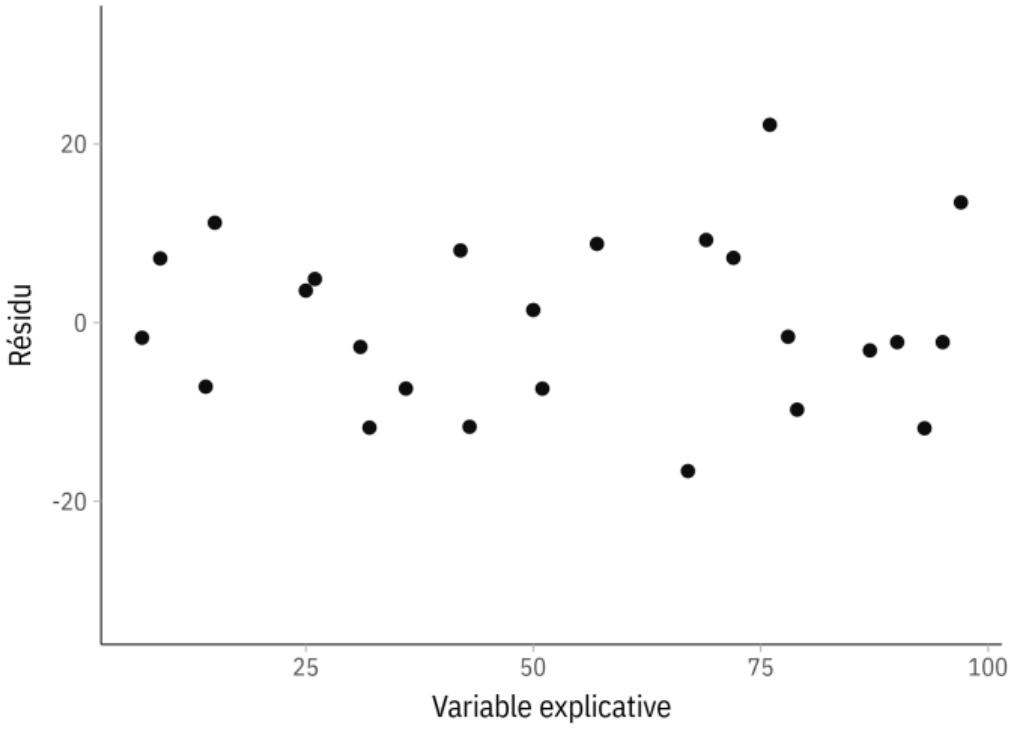
# Résidus



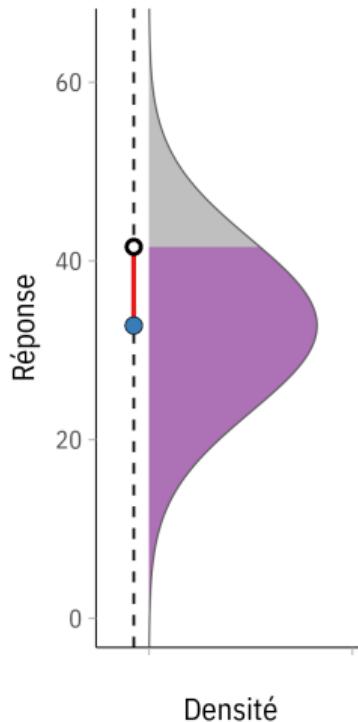
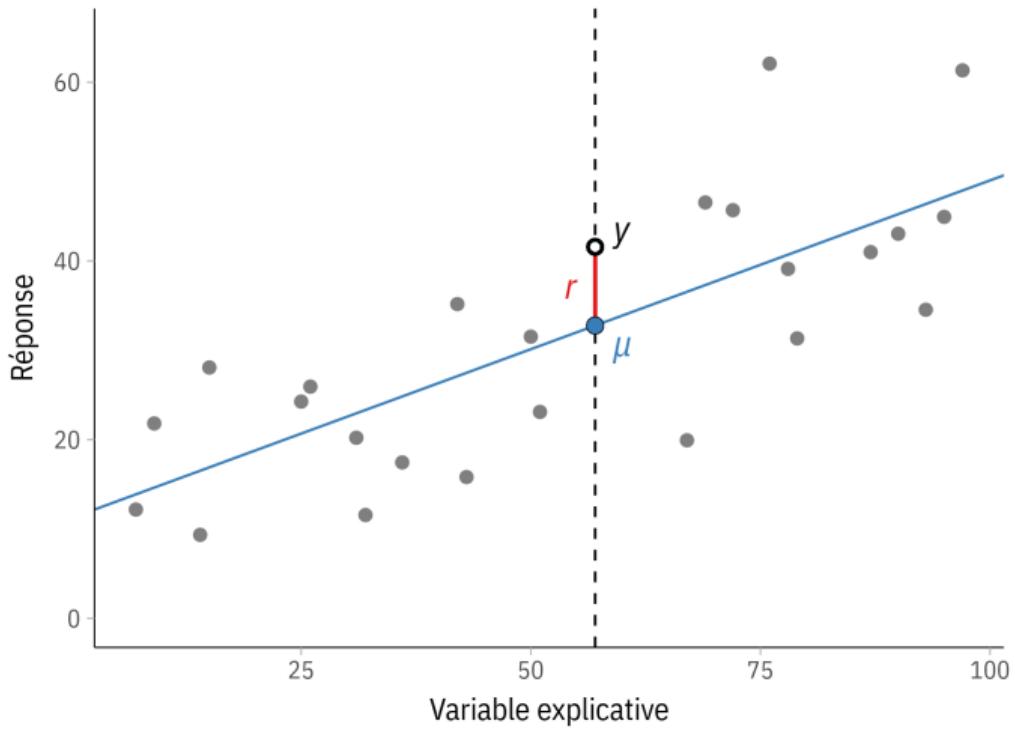
# Résidus



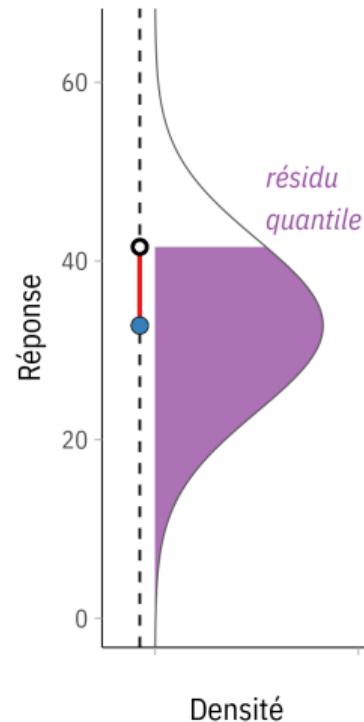
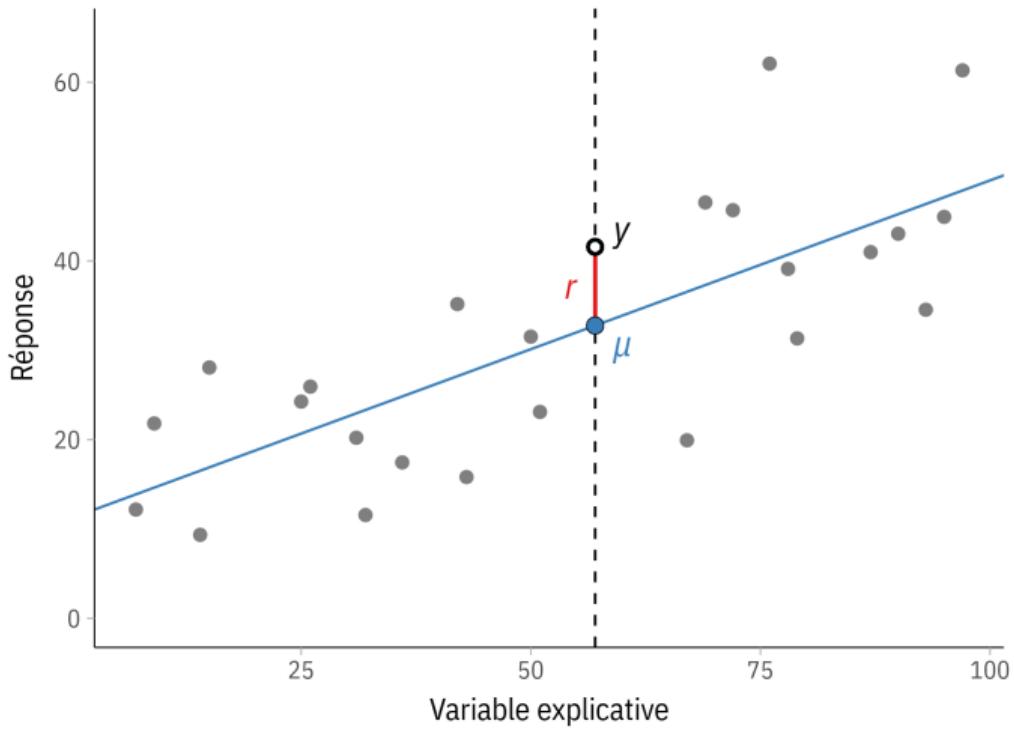
# Résidus



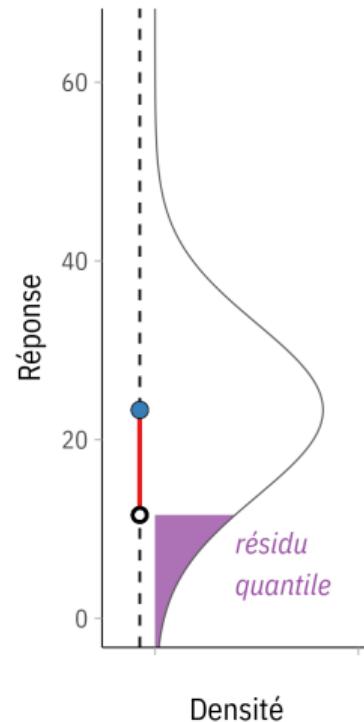
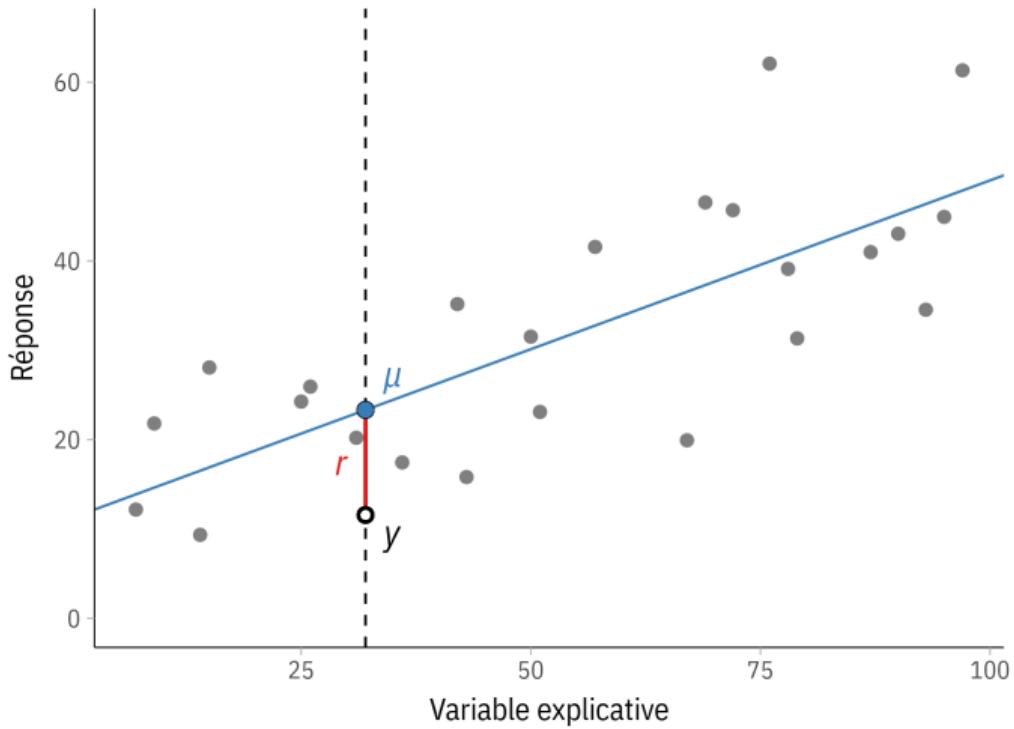
# Résidus quantiles



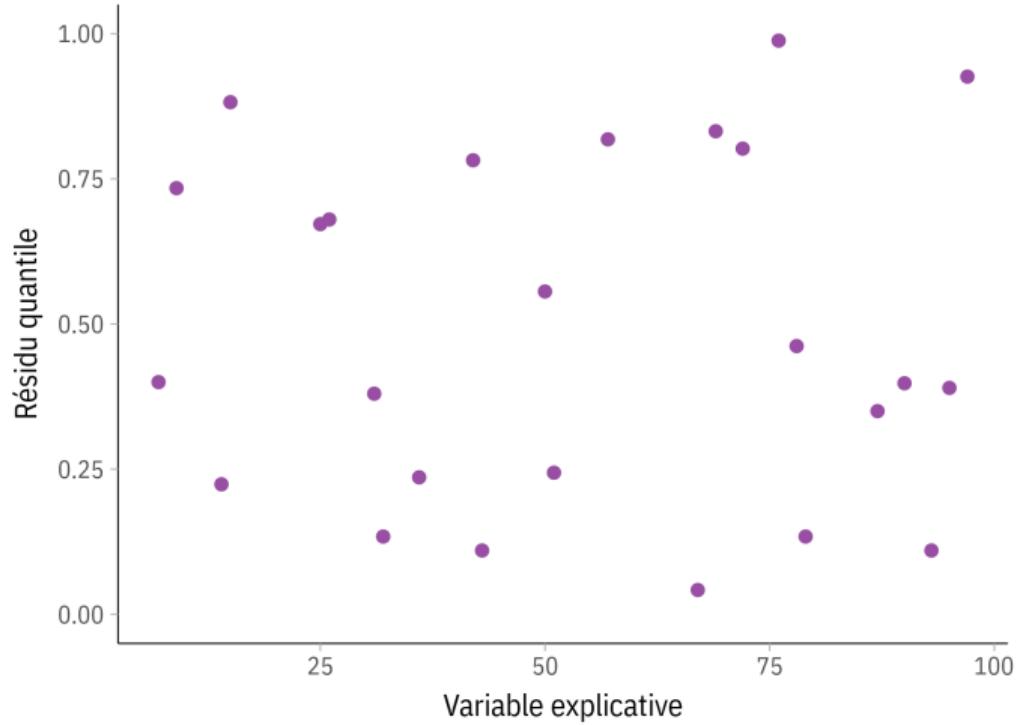
# Résidus quantiles



# Résidus quantiles



# Résidus quantiles



# Modèle linéaire

Pour chaque observation  $i$ :

$$y_i \sim \mathcal{N}(\mu_i, \sigma^2)$$

# Modèle linéaire

Pour chaque observation  $i$ :

$$y_i \sim \mathcal{N}(\mu_i, \sigma^2)$$

$$\mu_i = \alpha + \beta x_i + \cdots$$

# Modèle linéaire

Pour chaque observation  $i$ :

Résidu:  $r_i \sim \mathcal{N}(0, \sigma^2)$

$$y_i \sim \mathcal{N}(\mu_i, \sigma^2)$$

$$\mu_i = \alpha + \beta x_i + \cdots$$

$$y_i = \mu_i + r_i$$

# Modèle linéaire

Pour chaque observation  $i$ :

$$y_i \sim \mathcal{N}(\mu_i, \sigma^2)$$

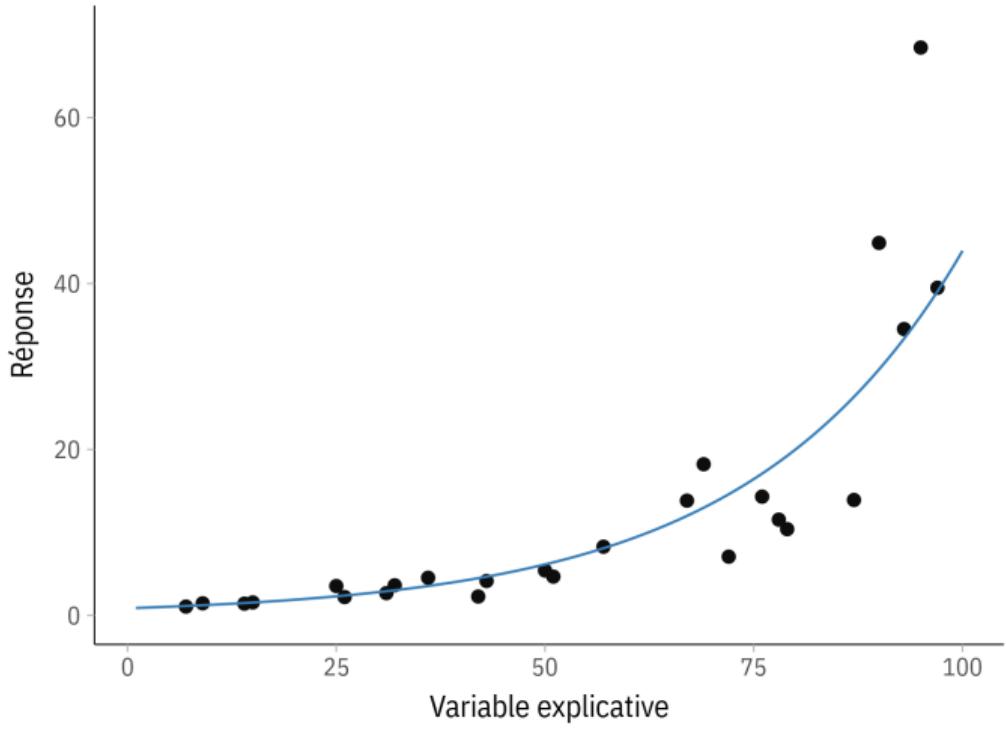
$$\mu_i = \alpha + \beta x_i + \cdots$$

$$y_i = \mu_i + r_i$$

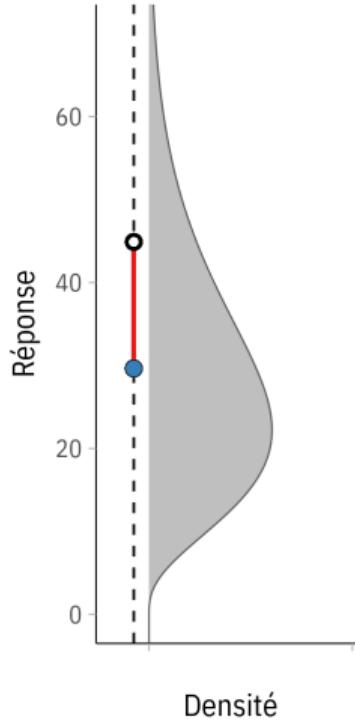
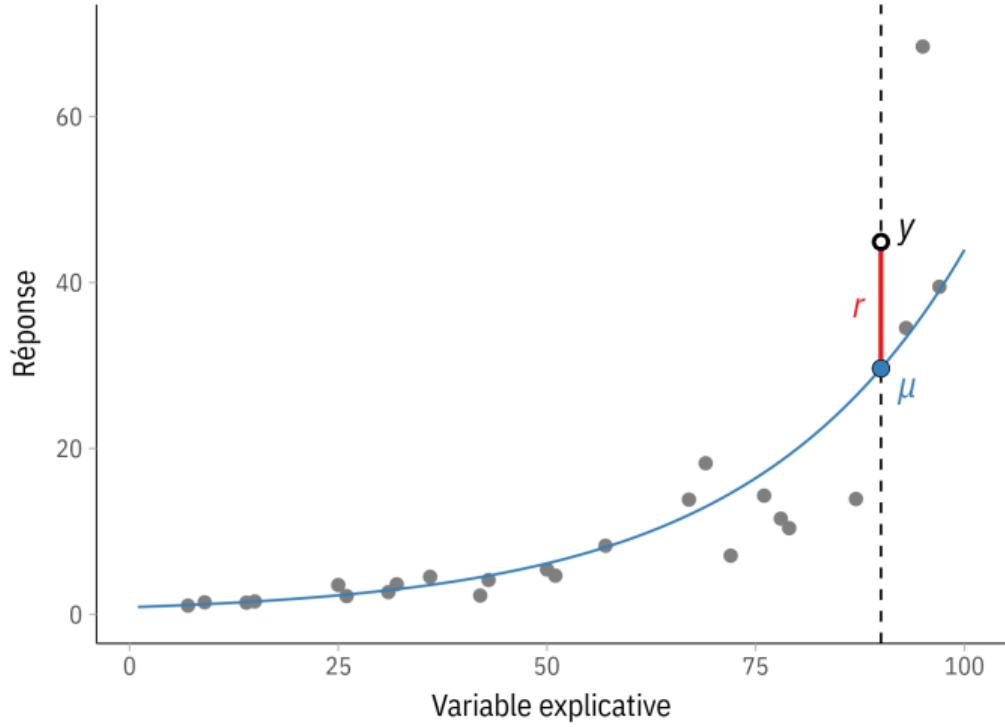
Résidu:  $r_i \sim \mathcal{N}(0, \sigma^2)$

Résidu quantile:  $r_{Qi} \sim \mathcal{U}_{[0,1]}$

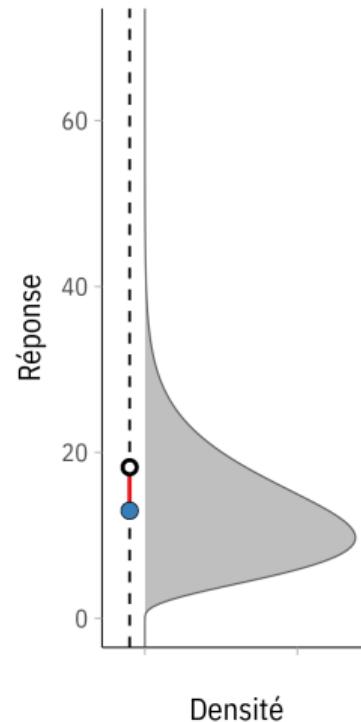
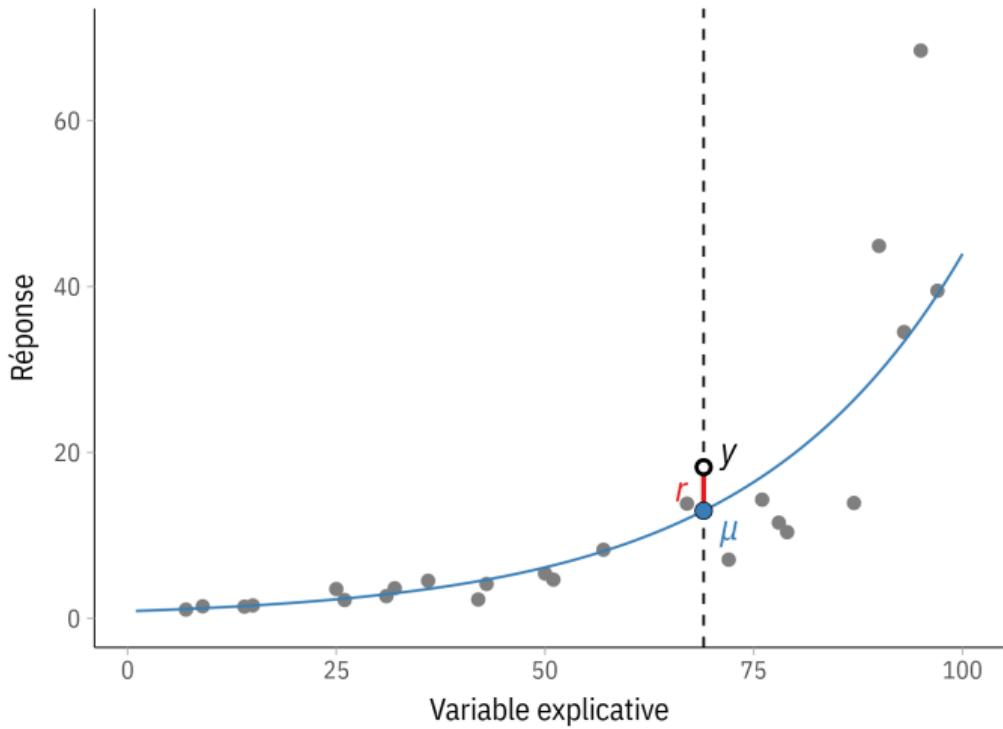
# Résidus du modèle linéaire généralisé



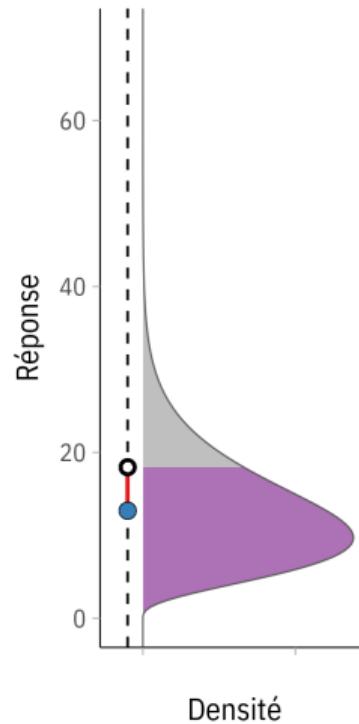
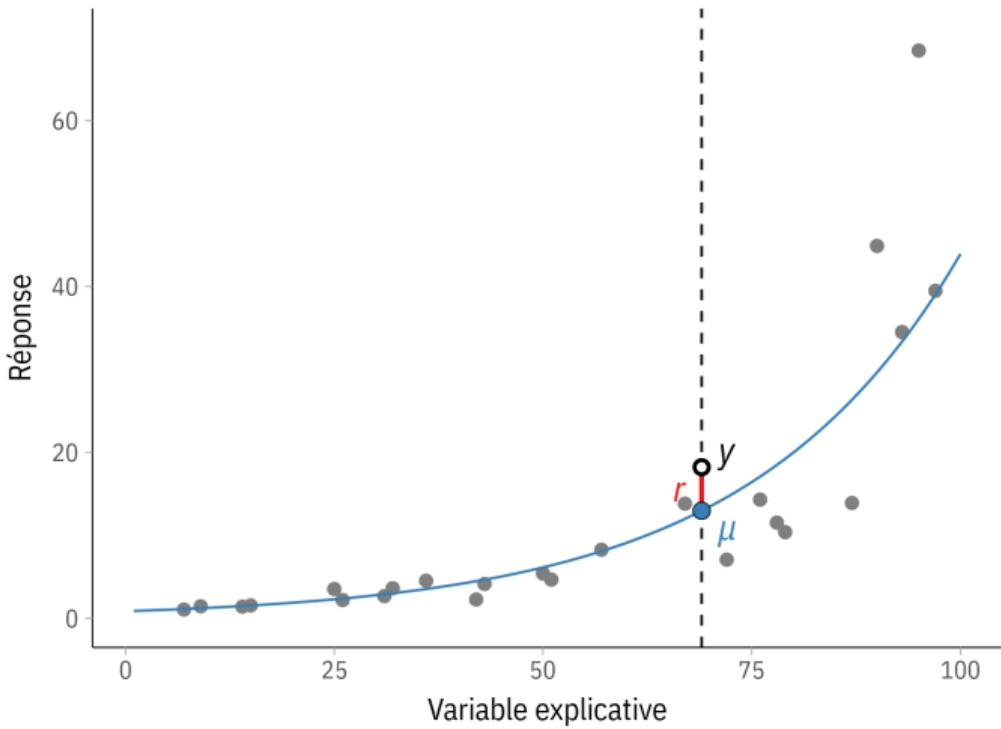
# Résidus du modèle linéaire généralisé



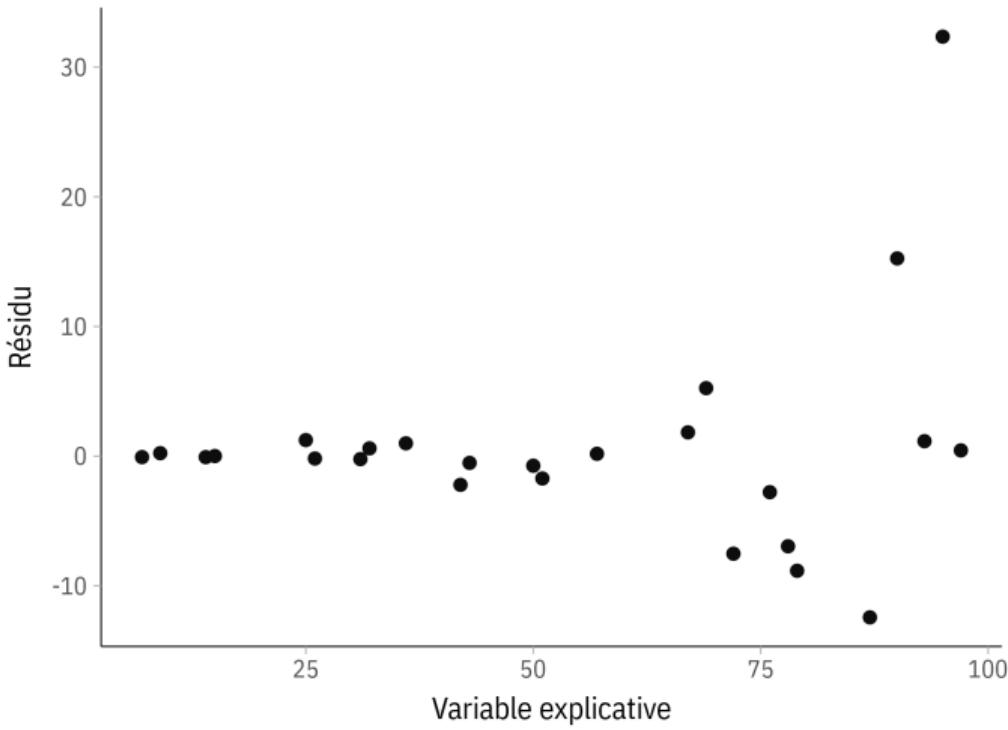
# Résidus du modèle linéaire généralisé



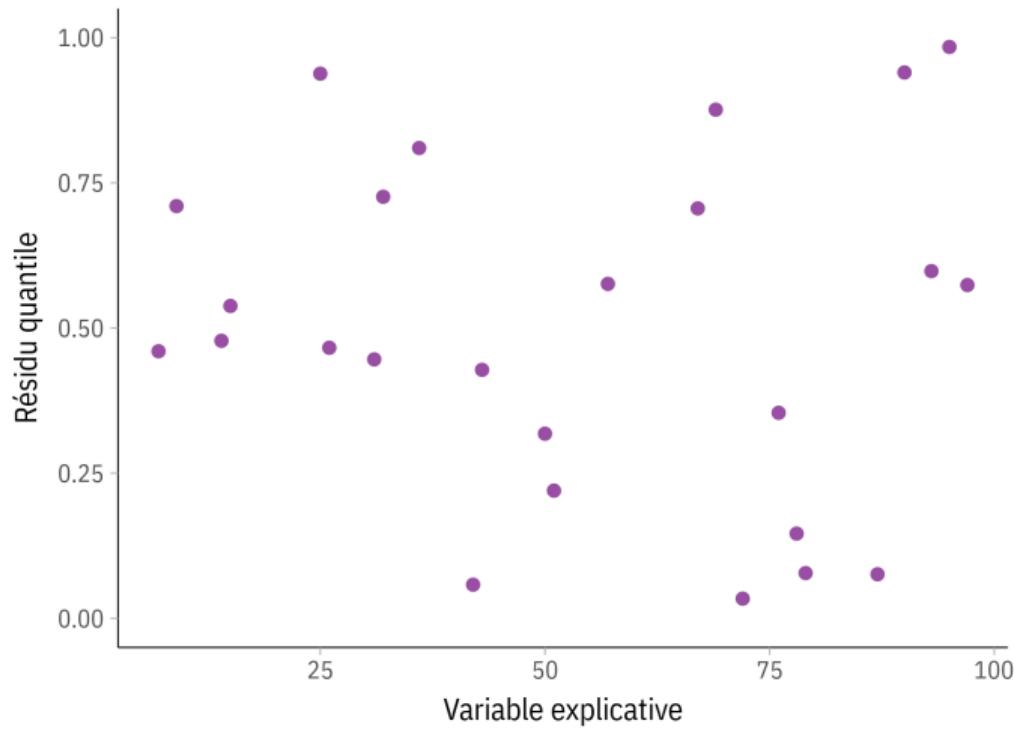
# Résidus du modèle linéaire généralisé



# Résidus du modèle linéaire généralisé



# Résidus du modèle linéaire généralisé



# Modèle linéaire généralisé

Pour chaque observation  $i$ :

$$y_i \sim \text{EF}(\mu_i, \phi)$$

$$g(\mu_i) = \alpha + \beta x_i + \cdots$$

$$y_i = \mu_i + r_i$$

Résidu:  $r_i \sim ?$

Résidu quantile:  $r_{Qi} \sim \mathcal{U}_{[0,1]}$

# Modèle linéaire généralisé

Pour chaque observation  $i$ :

$$y_i \sim \text{EF}(\mu_i, \phi)$$

$$g(\mu_i) = \alpha + \beta x_i + \cdots$$

$$y_i = \mu_i + r_i$$

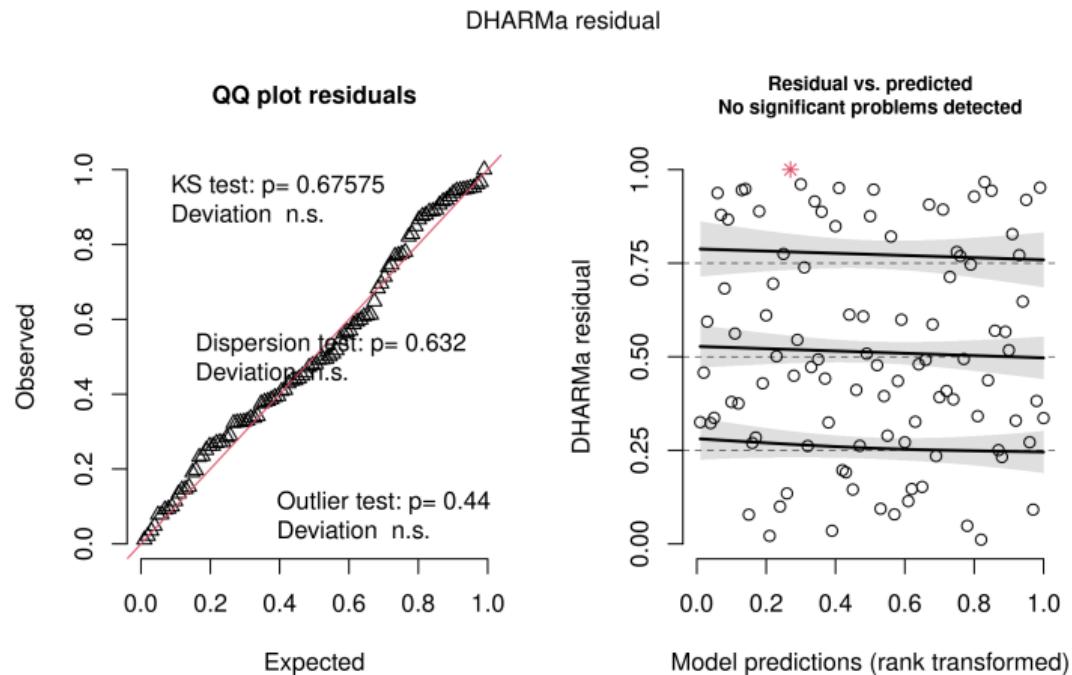
Résidu:  $r_i \sim ?$

Résidu quantile:  $r_{Qi} \sim \mathcal{U}_{[0,1]}$

Il existe d'autres types de résidus (par exemple *Pearson* où *déviance*) qui suivent une distribution presque normale *sous certaines conditions*.

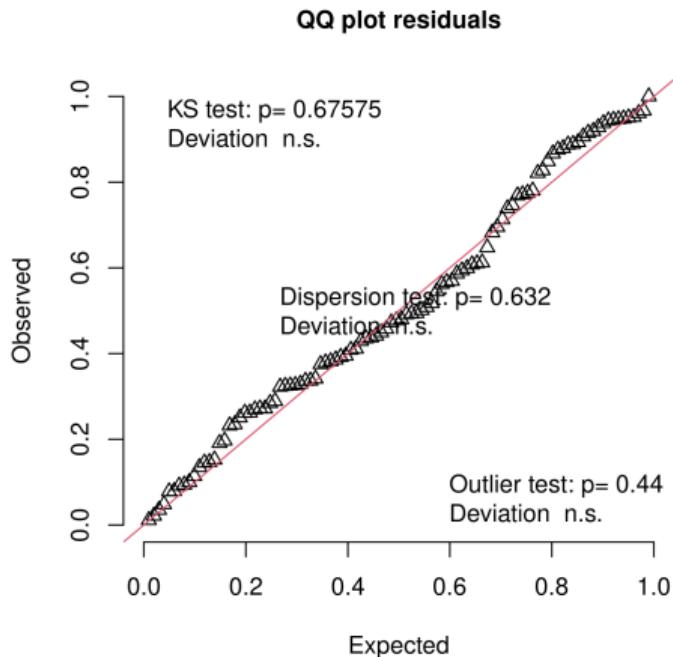
## Graphiques de résidus quantiles

# Graphiques de résidus quantiles

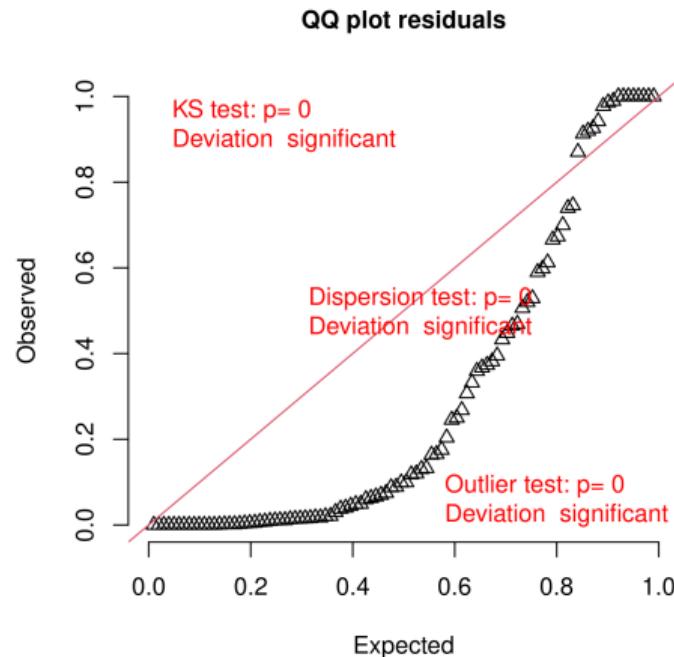


# Graphiques de résidus quantiles

Dispersion correcte

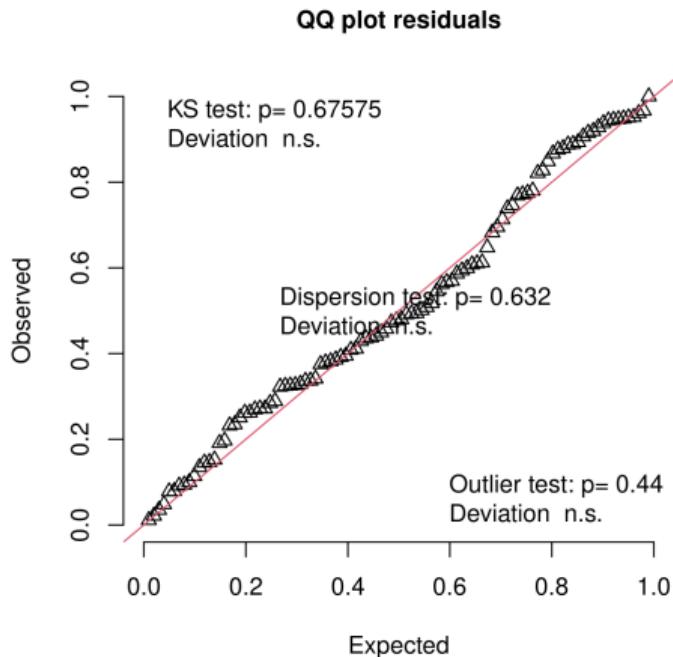


Surdispersion

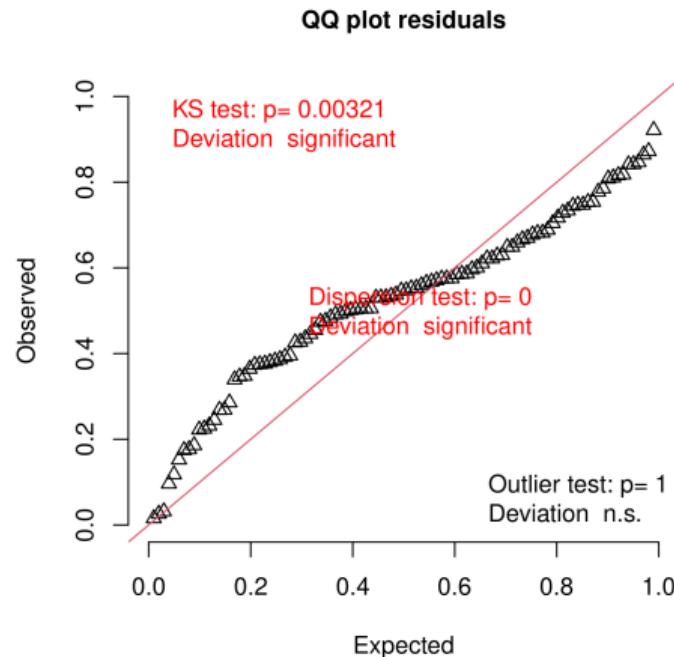


# Graphiques de résidus quantiles

Dispersion correcte



Sousdispersion



# **Surdispersion et sousdispersion**

## **Surdispersion (*overdispersion*)**

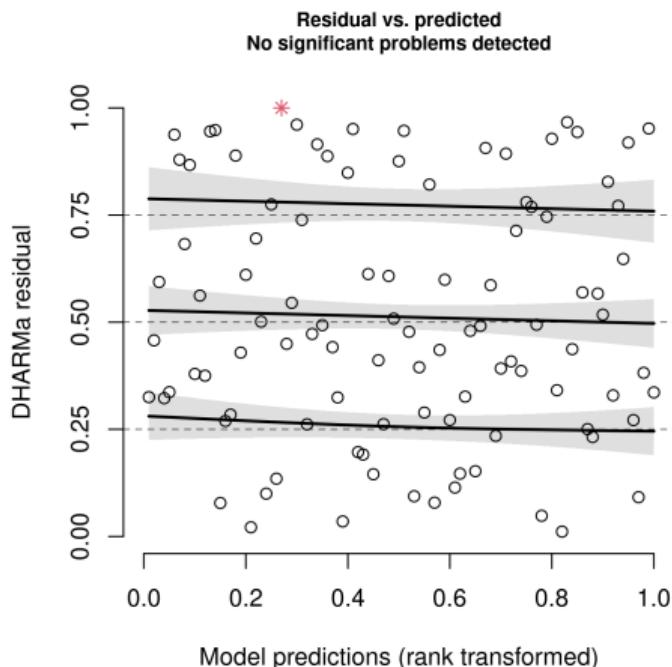
- Il y a plus de résidus extrêmes que prévu par le modèle.
- Le modèle ne prend pas en compte une partie de l'incertitude.
- Les intervalles de confiance sont trop petits et les valeurs p sont trop faibles.

## **Sousdispersion (*underdispersion*)**

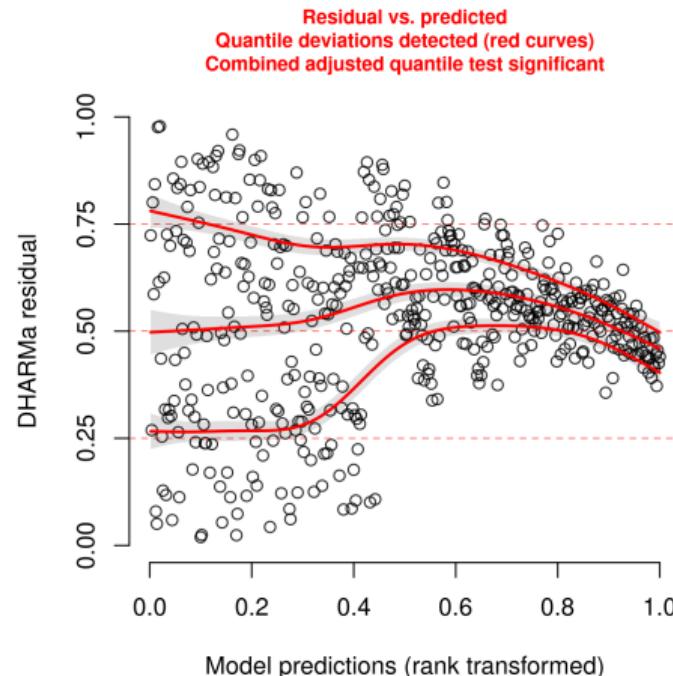
- Il y a moins de résidus extrêmes que prévu par le modèle.
- Le modèle prévoit trop de incertitude.
- Souvent moins grave que la surdispersion parce que l'inférence devient plus conservatrice.

# Graphiques de résidus quantiles

## Homoscédasticité



## Hétéroscédaстicité



# Rapporter la validation par résidus quantiles

## La méthode

Dunn, P. K., & Smyth, G. K. (1996). Randomized Quantile Residuals. *Journal of Computational and Graphical Statistics*, 5(3), 236–244.  
<https://doi.org/10.1080/10618600.1996.10474708>

## Le package

Hartig, F. (2022). DHARMA : Residual diagnostics for hierarchical (multi-level / mixed) regression models. R package version 0.4.6,  
<https://CRAN.R-project.org/package=DHARMA>

# Rapporter la validation par résidus quantiles

## Comment citer

Model assumptions were validated by inspecting quantile residuals (Dunn & Smyth, 1996), as implemented in the *R* package DHARMA (Hartig, 2022).

# Rapporter la validation par résidus quantiles

## Comment citer

Model assumptions were validated by inspecting quantile residuals (Dunn & Smyth, 1996), as implemented in the *R* package DHARMA (Hartig, 2022).

Il est recommandé d'inclure les graphiques résiduels dans l'annexe ou dans le matériel supplémentaire.

**Quelques distributions pour la variable de réponse**

# Quelques distributions pour la variable de réponse

Distribution	Support	Commentaire
<b>Normale</b> (Gaussien)	$(-\infty, \infty)$	Variance homogène pour toutes les observations.

# Quelques distributions pour la variable de réponse

Distribution	Support	Commentaire
<b>Normale</b> (Gaussien)	$(-\infty, \infty)$	Variance homogène pour toutes les observations.
<b>Gamma</b>	$(0, \infty)$	Variance égale à la moyenne au carré.

# Quelques distributions pour la variable de réponse

Distribution	Support	Commentaire
<b>Normale</b> (Gaussien)	$(-\infty, \infty)$	Variance homogène pour toutes les observations.
<b>Gamma</b>	$(0, \infty)$	Variance égale à la moyenne au carré.
<b>Poisson</b>	$\{0, 1, 2, \dots\}$	Variance égale à la moyenne. Souvent problèmes de surdispersion.

# Quelques distributions pour la variable de réponse

Distribution	Support	Commentaire
<b>Normale</b> (Gaussien)	$(-\infty, \infty)$	Variance homogène pour toutes les observations.
<b>Gamma</b>	$(0, \infty)$	Variance égale à la moyenne au carré.
<b>Poisson</b>	$\{0, 1, 2, \dots\}$	Variance égale à la moyenne. Souvent problèmes de surdispersion.
<b>Binomiale négative</b>	$\{0, 1, 2, \dots\}$	Variance augmente avec la moyenne (mais plus fortement que pour la distribution Poisson).

# Quelques distributions pour la variable de réponse

Distribution	Support	Commentaire
<b>Normale</b> (Gaussien)	$(-\infty, \infty)$	Variance homogène pour toutes les observations.
<b>Gamma</b>	$(0, \infty)$	Variance égale à la moyenne au carré.
<b>Poisson</b>	$\{0, 1, 2, \dots\}$	Variance égale à la moyenne. Souvent problèmes de surdispersion.
<b>Binomiale négative</b>	$\{0, 1, 2, \dots\}$	Variance augmente avec la moyenne (mais plus fortement que pour la distribution Poisson).
<b>Tweedie</b> (avec $0 < \xi < 1$ )	$[0, \infty)$	Variance augmente avec la moyenne (mais d'une façon plus flexible que les distributions Gamma et Poisson). Permet de prendre en compte un excès de zéros.

# Quelques distributions pour la variable de réponse

---

Distribution	Support	Commentaire
<b>Normale</b> (Gaussien)	$(-\infty, \infty)$	Variance homogène pour toutes les observations.
<b>Gamma</b>	$(0, \infty)$	Variance égale à la moyenne au carré.
<b>Poisson</b>	$\{0, 1, 2, \dots\}$	Variance égale à la moyenne. Souvent problèmes de surdispersion.
<b>Binomiale négative</b>	$\{0, 1, 2, \dots\}$	Variance augmente avec la moyenne (mais plus fortement que pour la distribution Poisson).
<b>Tweedie</b> (avec $0 < \xi < 1$ )	$[0, \infty)$	Variance augmente avec la moyenne (mais d'une façon plus flexible que les distributions Gamma et Poisson). Permet de prendre en compte un excès de zéros.
<b>Bernoulli</b>	$\{0, 1\}$	Agrégation nécessaire pour détecter une possible surdispersion. Équivalente à une distribution binomiale avec un essai.

---

## **Plus d'information sur les modèles généralisés**

Zuur, A. F., Ieno, E. N., Walker, N., Saveliev, A. A., & Smith, G. M. (2009). Mixed effects models and extensions in ecology with R. Springer New York.  
<https://doi.org/10.1007/978-0-387-87458-6>

Dormann, C. (2020). Environmental Data Analysis : An Introduction with Examples in R. Springer International Publishing. <https://doi.org/10.1007/978-3-030-55020-2>

## **Exemple 2: Effet de l'ozone sur les semis de l'épinette de Sitka**

## Exemple 2: Effet de l'ozone sur les semis de l'épinette de Sitka



**Données**

`sitka.rds`

**Références**

Diggle, P., Heagerty, P., Liang, K.-Y., & Zeger, S. (2002). Analysis of Longitudinal Data (Second Edition). Oxford University Press.

Données recueillies par Dr Peter Lucas (*Biological Sciences Division, Lancaster University*).

Image: Wikimedia (Brandon Kuschel)

## Exemple 2: Effet de l'ozone sur les semis de l'épinette de Sitka



Données

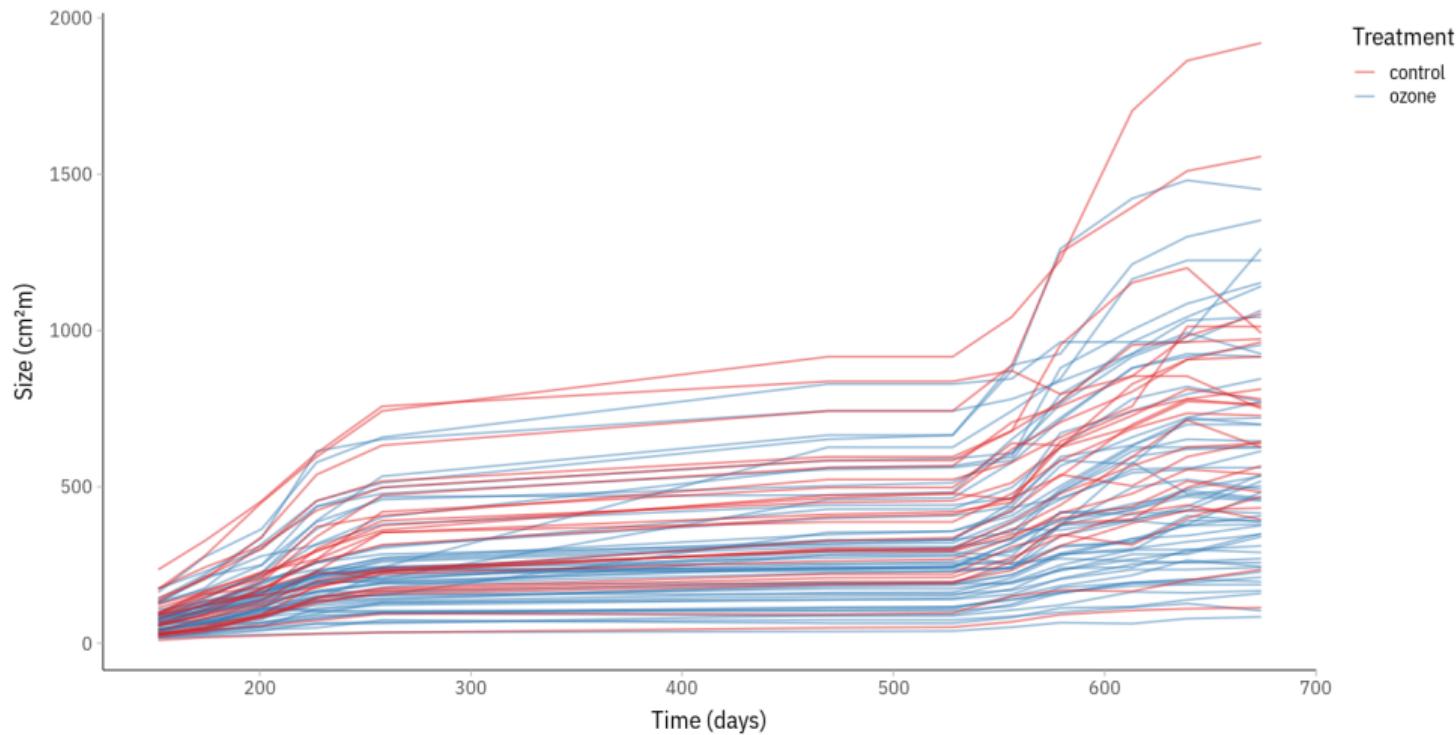
`sitka.csv`

Variables

- `tree.id` : identité de l'arbre (79 individus) ;
- `day` : nombre de jours depuis le 1er janvier 1988 ;
- `size` : taille de l'arbre (hauteur multipliée par le diamètre,  $10^{-4} \text{ m}^3$ ) ;
- `treatment` : indique si les arbres sont maintenus dans un environnement normal (`control`) ou enrichi ( $70 \text{ nl l}^{-1}$ ) en ozone (`ozone`).

Image: Wikimedia (Brandon Kuschel)

## Exemple 2: Effet de l'ozone sur les semis de l'épinette de Sitka



## **Exemple 3: Distribution des lichens en Suède**

# Exemple 3: Distribution des lichens en Suède



## Données

lichen.csv

## Références

Esseen, Per-Anders et al. (2022), Multiple drivers of large-scale lichen decline in boreal forest canopies, Dryad, Dataset, <https://doi.org/10.5061/dryad.2ngf1vhq5>

Esseen, P.-A., Ekström, M., Grafström, A., Jonsson, B. G., Palmqvist, K., Westerlund, B., & Ståhl, G. (2022). Multiple drivers of large-scale lichen decline in boreal forest canopies. *Global Change Biology*, 28(10), 3293–3309. <https://doi.org/10.1111/gcb.16128>

# Exemple 3: Distribution des lichens en Suède



Données

lichen.rds

Variables

- `tree.id` : identité unique de l'arbre ;
- *Informations sur l'inventaire* : `ip` (période d'inventaire, 1 ou 2) ; `region` (région de l'inventaire) ; `year` (année de l'évaluation) ;
- `east, north` : coordonnées projetées dans la grille de référence suédoise (EPSG : 3006) ;
- `species` : genre du lichen (soit `Usnea`, `Aleactoria`, ou `Bryoria`) ;
- `occurrence` : indique si des lichens du genre correspondant ont été trouvés sur l'arbre (1) ou pas (0) ;

# Exemple 3: Distribution des lichens en Suède



## Données

lichen.rds

## Variables (suite)

- ...
- **Variables environnementales** : `mat` (proportion de forêts matures dans un rayon de 100 m) ; `temp` (température annuelle moyenne, °C) ; `rain` (cumul annuel de pluie, mm) ; `ndep` (dépôt annuel moyen d'azote, kg N ha<sup>-1</sup> an<sup>-1</sup>) ;
- **Variables au niveau de l'arbre** : `dbh` (diamètre à hauteur de poitrine, mm) ; `crl` (limite du houppier, m) ;
- **Variables au niveau du peuplement** : `bas` (surface terrière, m<sup>2</sup> ha<sup>-1</sup>), `age` (âge du peuplement) ;

# Exemple 3: Distribution des lichens en Suède



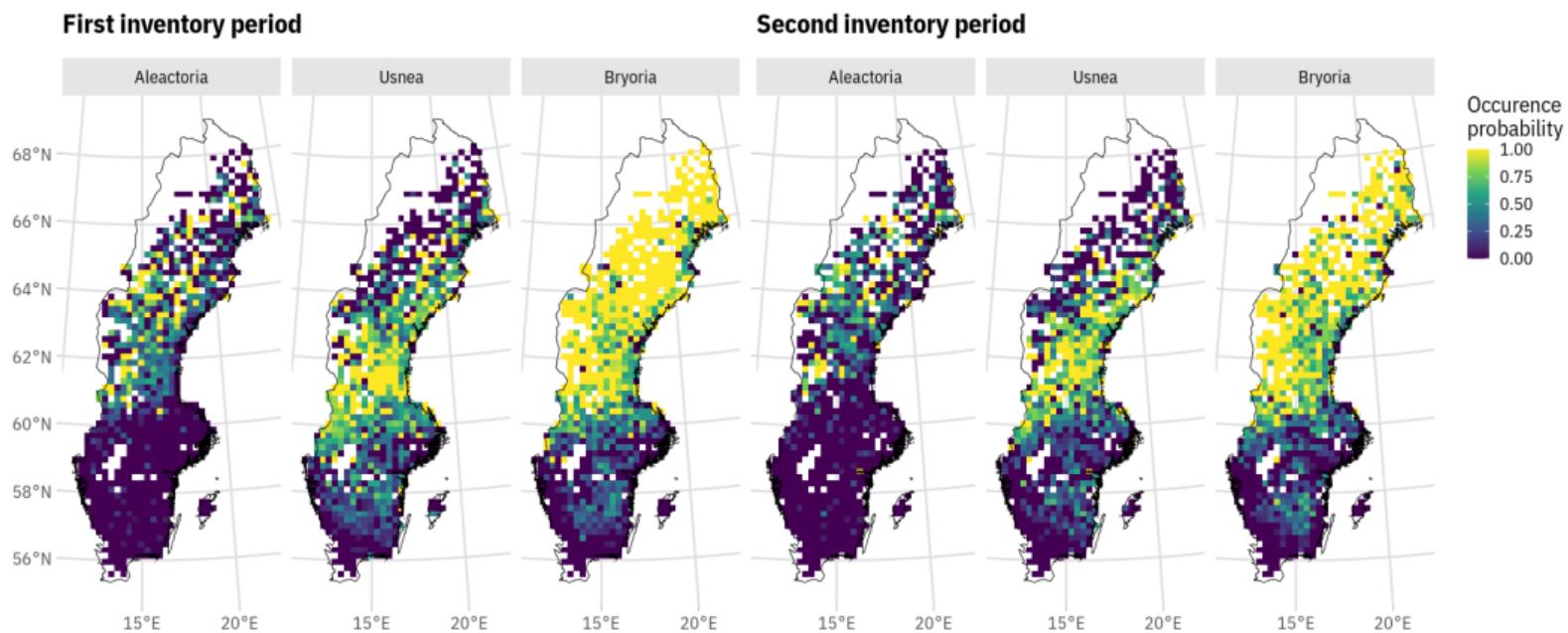
Données

`lichen.rds`

Variables (suite)

- ...
- Grille d'agrégation spatiale (20km×20km) : `east.agg`, `north.agg`, `rast.agg.id`

# Exemple 3: Distribution des lichens en Suède



## **Fonctions importantes**

# Fonctions importantes

## Fonctions principales

---

Fonction	Déscription
<code>simulateResiduals()</code>	Simuler de résidus quantiles randomisés.
<code>plot()</code>	Produire un graphique QQ (comme <code>testUniformity()</code> ) et un graphique des résidus en fonction des valeurs prédictes (comme <code>testQuantiles()</code> ).
<code>plotResiduals()</code>	Produire graphique des résidus en fonction des valeurs prédictes ou d'une autre variable et valider l'absence de patrons et heteroscedasticité.

---

# Fonctions importantes

## Tests individuels (utilisés par les autres fonctions)

---

Fonction	Déscription
<code>testUniformity()</code>	Produire un graphique QQ et vérifier l'uniformité de la distribution des résidus.
<code>testQuantile()</code>	Produire un graphique des résidus en fonction des valeurs prédictes et vérifier l'absence de patrons et heteroscedasticité.
<code>testDispersion()</code>	Vérifier la dispersion.
<code>testOutliers()</code>	Vérifier la fréquence des résidus extrêmes.

---

# Fonctions importantes

## Autocorrelation temporelle et spatiale

---

Fonction	Déscription
<code>recalculateResiduals()</code>	Agréger les résidus par groupe (par exemple par période de temps ou par lieu géographique).
<code>testTemporalAutocorrelation()</code>	Vérifier l'autocorrelation temporelle résiduelle.
<code>testSpatialAutocorrelation()</code>	Vérifier l'autocorrelation spatiale résiduelle.

---

# Fonctions importantes

## Excès de zéros

---

Fonction	Déscription
<code>testZeroInflation()</code>	Vérifier s'il y a plus de zéros exacts que ceux pris en compte par le modèle.

---

# Fonctions importantes

## Modèles bayésiens et autres

---

Fonction	Déscription
<code>createDHARMA()</code>	Calculer les résidus à partir d'un ensemble de simulations. Utile pour valider des modèles bayésiens.

---

*Voir aussi:*

<https://cran.r-project.org/web/packages/DHARMA/vignettes/DHARMAForBayesians.html>