Modèles linéaires mixtes, modèles linéaires généralisés mixtes (réponses binaires), analyse de données de durées.

Frédérique Letué

Modèles linéaires mixtes, modèles linéaires généralisés mixtes, analyse de données de durées

Frédérique Letué

Univ. Grenoble Alpes, LJK, F-38000 Grenoble, France
CNRS, Laboratoire Jean Kuntzmann, UMR CNRS 5224, F-38000 Grenoble, France

28 juin 2017

Collaborateurs

Tous les exemples présentés dans la formation sont issus de travaux en collaboration avec

- des statisticiennes du LJK [1] : Adeline Leclercq-Samson, Marie-José Martinez ;
- des chercheurs en sciences du langage du GIPSA-lab [2] : Diane Caussade, Sandra Cornaz, Silvain Gerber, Nathalie Henrich-Bernardoni, Nathalie Vallée, Anne Vilain, Coriandre Vilain
- et du LIDILEM [3] : Jean-Marc Coletta

[1] Univ. Grenoble Alpes, LJK, F-38000 Grenoble, France
CNRS, Laboratoire Jean Kuntzmann, UMR CNRS 5224, F-38000 Grenoble, France
[2] Univ. Grenoble Alpes, GIPSA-Lab, F-38000 Grenoble, France
CNRS, GIPSA-Lab, F-38000 Grenoble, France
[3] Univ. Grenoble Alpes, LIDILEM, F-38000 Grenoble, France

Plan du chapitre

Généralités sur la modélisation en statistique

Une étude statistique a pour but de répondre à une question dans un domaine d'application particulier à partir d'un jeu de données.
Si les données sont issues d'une expérience, on les modélisera selon les étapes suivantes :

1. étape(s) modélisation : proposer un modèle probabiliste, en adéquation avec le type de données disponibles, duquel les données "pourraient" être issues ;
2. étape(s) de validation de modèle : est-il raisonnable de supposer que les données sont issues du modèle ?
   - si non, retour à l'étape de modélisation ;
   - si oui, étape suivante ;
3. étape d' interprétation : formulation mathématique des questions posées à partir des paramètres du modèle, tests statistiques adaptés, et réponses aux questions.

Jeu de données "production de voyelles non-natives"

Ce jeu de données est issu de la thèse de Sandra Cornaz [4]

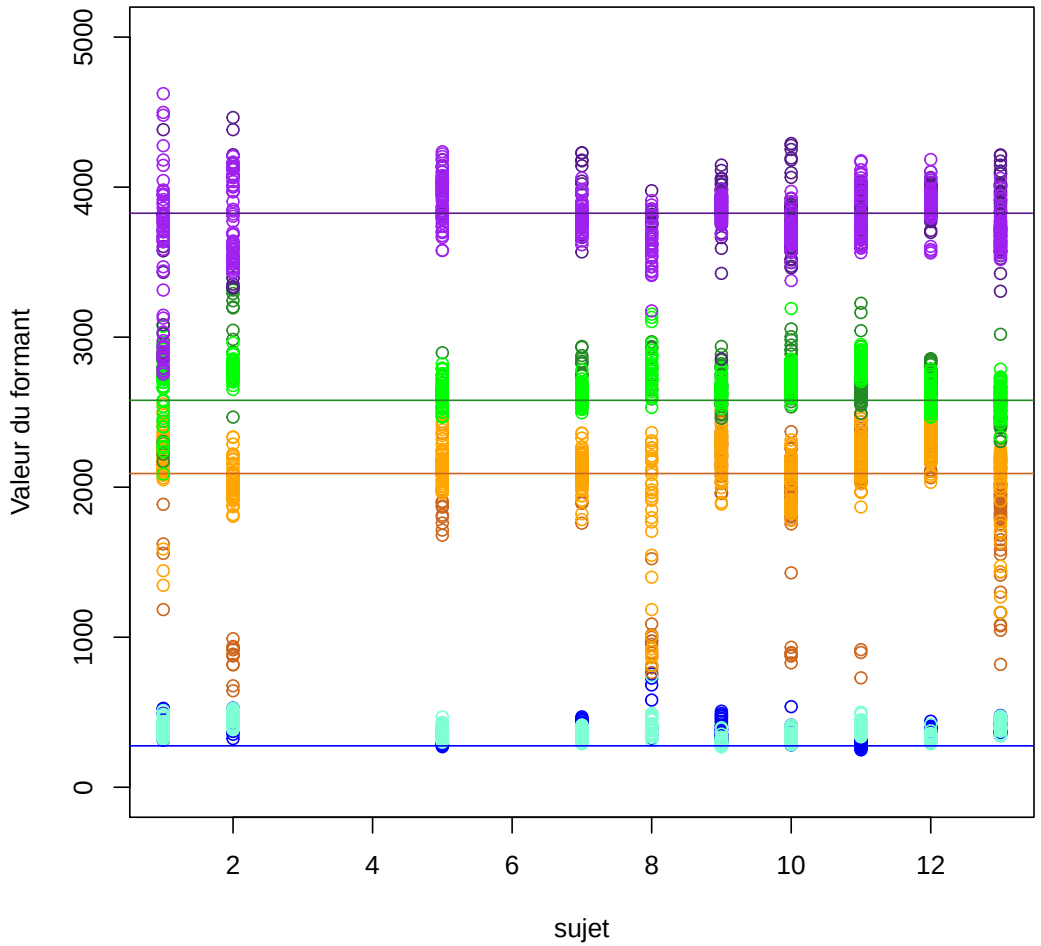
- 10 locutrices italiennes qui apprennent le français en langue étrangère,
- produisent des voyelles dont on mesure les hauteurs de formants (F1, F2, F3, F4),
- avant et après formation.

5 variables statistiques :

- hauteur de formant de la voyelle /y/ (F_{ftsk}) : variable d'intérêt (ou réponse), continue
- type de formant ($f = 1, \cdots, 4$) : facteur qualitatif
- test ($t \in \{\text{Avant, Après}\}$) : facteur qualitatif
- sujet ($s = 1, \cdots, 10$) : facteur qualitatif
- répétition ($k = 1, \cdots, n_s$) : variable qualitative

[4] [Cornaz] S. Cornaz Couffini (2014) L'apport de la voix chantée pour l'intégration phonético-phonologique d'une langue étrangère : application auprès d'italophones apprenants de FLE. *Thèse de l'Université Grenoble Alpes.*

Jeu de données "production de voyelles non-natives"



Jeu de données "production de voyelles non-natives"

Questions posées :

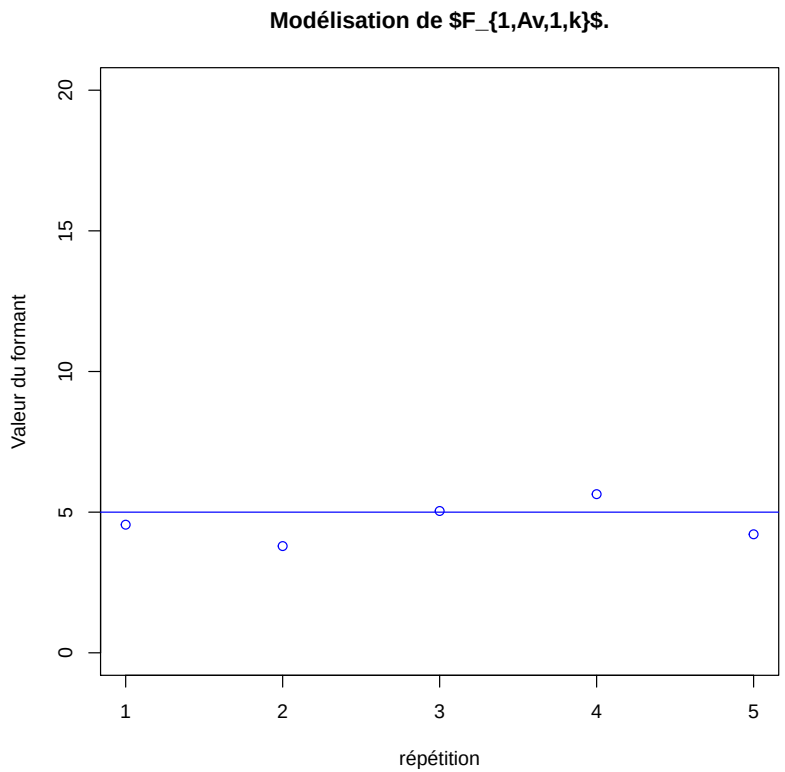
- Les formants sont-ils proches des valeurs de référence avant formation ?
- Jusqu'à quel point sont-ils améliorés après formation ?
- Les différences entre formants $(F_2 - F_1, F_3 - F_2, F_4 - F_3)$ sont-elles similaires avant et après formation ?

Modélisation

On s'intéresse d'abord aux mesures du formant F1, avant formation, pour le sujet 1.

$$F_{1,Av,1,k} = \mu + \varepsilon_{1,Av,1,k},$$

où $\varepsilon_{1,Av,1,k} \sim \mathcal{N}(0, \sigma^2)$.

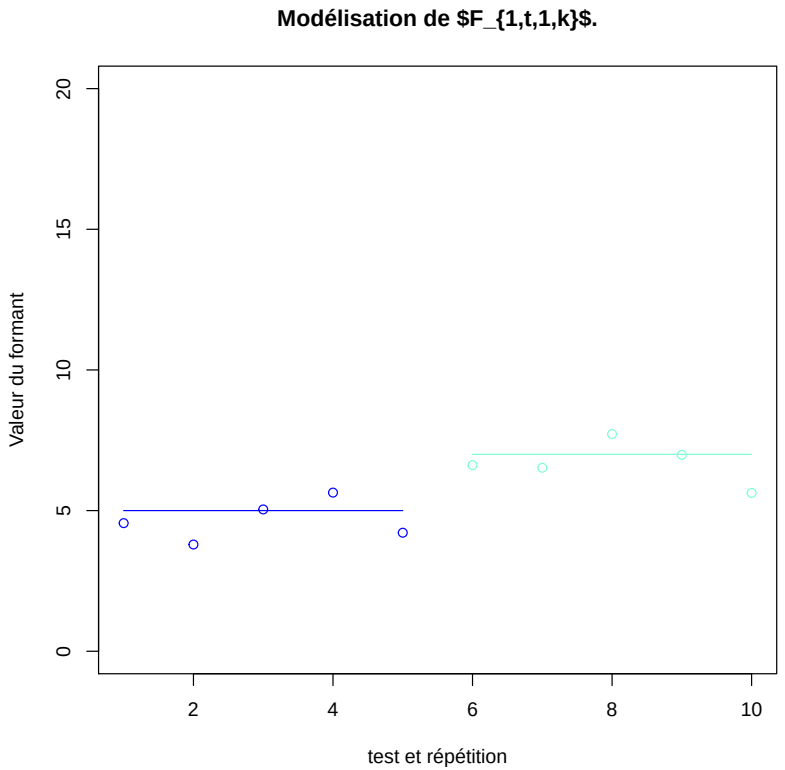


Modélisation

On ajoute ensuite l'effet fixe formation

$$F_{1,t,1,k} = \mu + \beta_t + \varepsilon_{1,t,1,k},$$

où $\varepsilon_{1,t,1,k} \sim \mathcal{N}(0, \sigma^2)$,
$\beta_{Av} = 0$

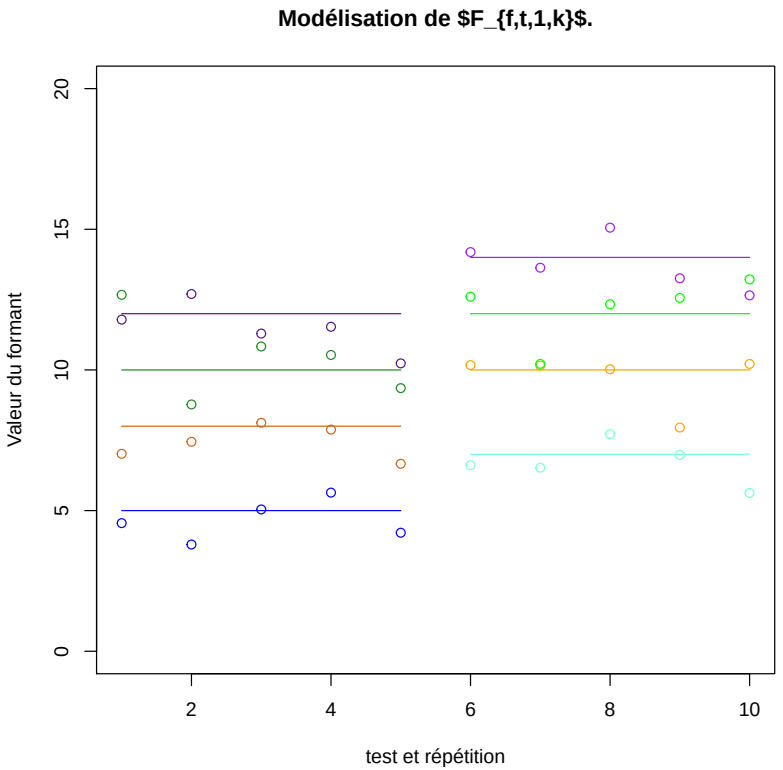


Modélisation

On ajoute encore l'effet fixe formant et ses interactions avec la formation

$$F_{f,t,1,k} = \mu + \alpha_f + \beta_t + \gamma_{ft} + \varepsilon_{f,t,1,k},$$

où $\varepsilon_{f,t,1,k} \sim \mathcal{N}(0, \sigma^2)$,
$\beta_{Av} = 0$,
$\alpha_1 = 0$,
$\gamma_{1,t} = \gamma_{f,Av} = 0$,

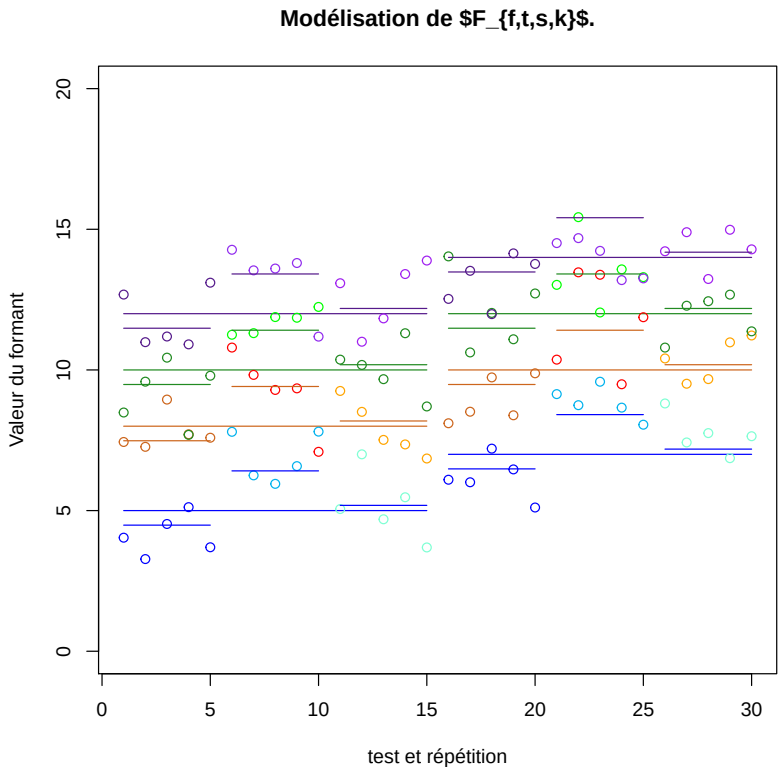


Modélisation

On ajoute enfin l'effet aléatoire sujet

$$F_{f,t,s,k} = \mu + \alpha_f + \beta_t + \gamma_{ft} + \xi_i + \varepsilon_{f,t,s,k},$$

où $\varepsilon_{f,t,s,k} \sim \mathcal{N}(0, \sigma^2)$,
$\beta_{Av} = 0$,
$\alpha_1 = 0$,
$\gamma_{1,t} = \gamma_{f,Av} = 0$,
$\xi_s \sim \mathcal{N}(0, \tau^2)$



Estimation des paramètres du modèle

A partir des données, on estime les paramètres du modèle :

$$\hat{\mu}, \hat{\alpha}_f, \hat{\beta}_t, \hat{\gamma}_{ft}, \hat{\sigma}, \hat{\tau},$$

et les effets aléatoires estimés : $\hat{\xi}_s$.
A partir de ces estimations, on peut calculer :

- des prédictions de populations : $\hat{\mu} + \hat{\alpha}_f + \hat{\beta}_t + \hat{\gamma}_{ft}$
- des prédictions individuelles : $\hat{F}_{fts} = \hat{\mu} + \hat{\alpha}_f + \hat{\beta}_t + \hat{\gamma}_{ft} + \hat{\xi}_s$
- des résidus : $\hat{\varepsilon}_{ftsk} = F_{f,t,s,k} - \hat{F}_{fts}$.

La qualité d'ajustement du modèle se mesure à l'aide des résidus.

Ajustement du modèle et analyse des résidus

On ajuste le modèle aux données et on examine les résidus

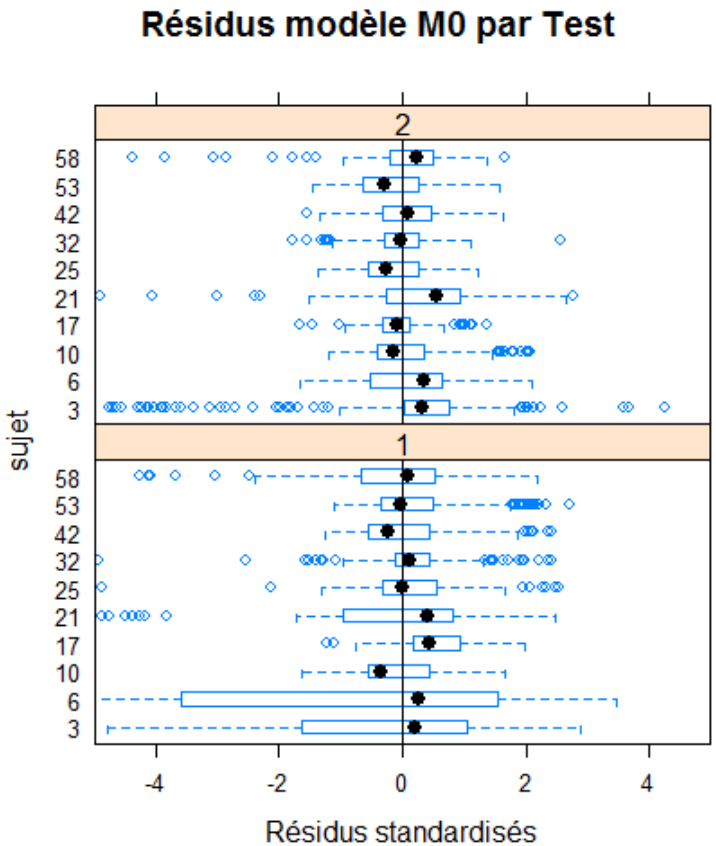


Figure: Boxplots des résidus standardisés par Test et sujet du modèle M_0.

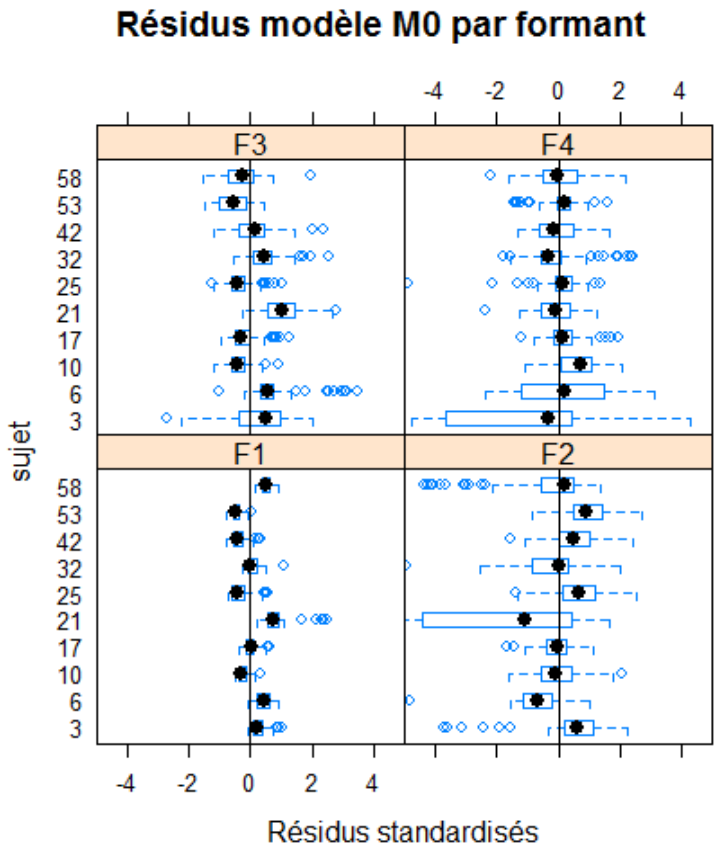


Figure: Boxplots des résidus standardisés par formant et sujet du modèle M_0.

Analyse du modèle M_0

On remarque que

- les résidus ne sont pas centrés par Test, ni par formant
- les résidus ont des dispersions similaires d'un Test à l'autre, mais pas d'un formant à l'autre.

On propose donc le modèle M_1 suivant :

$$F_{f,t,s,k} = \mu + \alpha_f + \beta_t + \gamma_{ft} + \xi_s + \xi_{fs} + \xi_{ts} + \varepsilon_{f,t,1,k},$$

où $\xi_{fs} \sim \mathcal{N}(0, \tau_1^2)$, et $\xi_{ts} \sim \mathcal{N}(0, \tau_2^2)$.

Ajustement du modèle $M1$ et analyse des résidus

On ajuste le modèle M_1 aux données et on examine les résidus

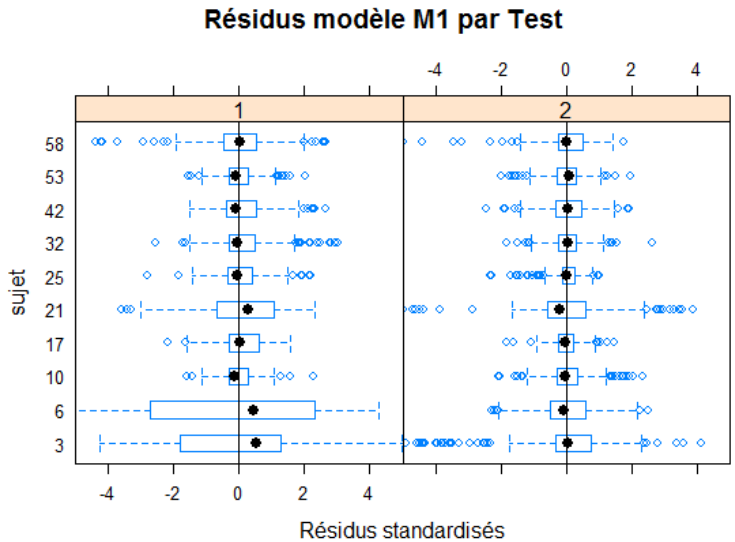


Figure: Boxplots des résidus standardisés par Test et sujet du modèle M_1.

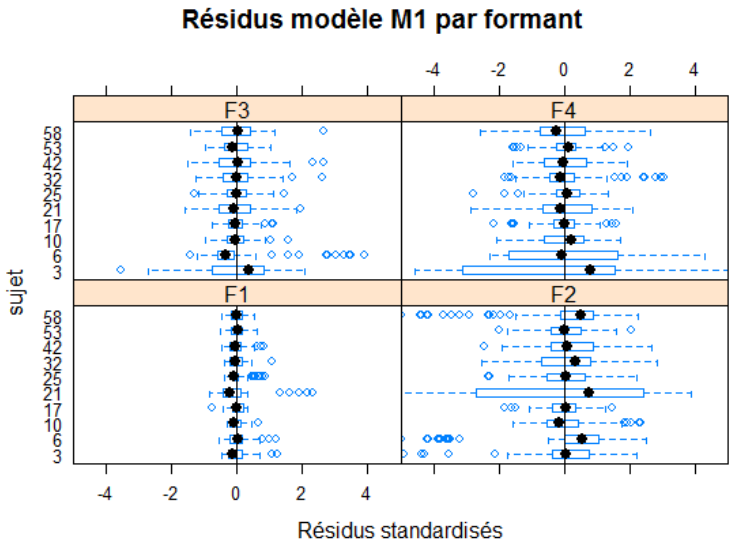


Figure: Boxplots des résidus standardisés par formant et sujet du modèle M_1.

Analyse du modèle M_1

On remarque que

- les résidus sont cette fois centrés par Test et par formant. Un test de comparaison de modèles confirme que le modèle M_1 ajuste mieux les données que le modèle M_0.
- les résidus ont des dispersions similaires d'un Test à l'autre, mais pas d'un formant à l'autre.

On propose donc le modèle M_2 suivant :

$$F_{f,t,s,k} = \mu + \alpha_f + \beta_t + \gamma_{ft} + \xi_s + \xi_{fs} + \xi_{ts} + \varepsilon_{f,t,1,k},$$

où $\varepsilon_{f,t,s,k} \sim \mathcal{N}(0, \sigma_f^2)$.

Ajustement du modèle $M2$ et analyse des résidus

On ajuste le modèle M_2 aux données et on examine les résidus

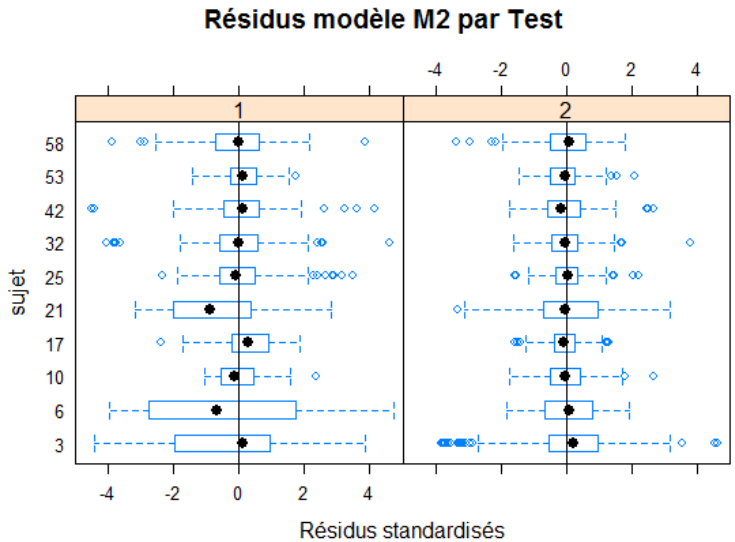


Figure: Boxplots des résidus standardisés par Test et sujet du modèle M_2.

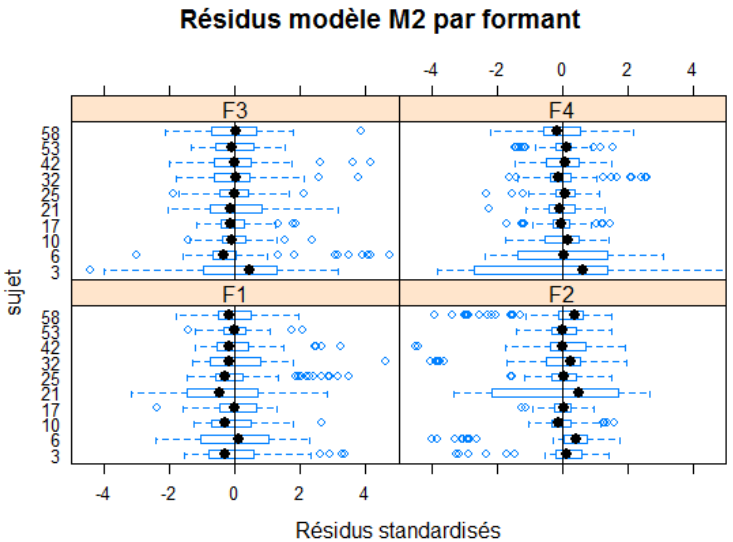


Figure: Boxplots des résidus standardisés par formant et sujet du modèle M_2.

Analyse du modèle M_2

On remarque que

- les résidus sont toujours centrés par Test et par formant.
- les résidus ont des dispersions similaires d'un Test à l'autre, et d'un formant à l'autre. Un test de comparaison de modèles confirme que le modèle M_2 ajuste mieux les données que le modèle M_1.

Pour l'instant, on a supposé toutes les données indépendantes sachant le sujet. Comment peut-on modéliser la dépendance entre formants ?

Ecriture matricielle du modèle M_2

$$\begin{bmatrix} F_{1tsk} \\ F_{2tsk} \\ F_{3tsk} \\ F_{4tsk} \end{bmatrix} = \mu \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} + \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \alpha_4 \end{bmatrix} + \beta_t \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} + \begin{bmatrix} \gamma_{1t} \\ \gamma_{2t} \\ \gamma_{3t} \\ \gamma_{4t} \end{bmatrix}$$

$$+\xi_s \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} + \begin{bmatrix} \xi_{1s} \\ \xi_{2s} \\ \xi_{3s} \\ \xi_{4s} \end{bmatrix} + \xi_{ts} \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} + \begin{bmatrix} \varepsilon_{1tsk} \\ \varepsilon_{2tsk} \\ \varepsilon_{3tsk} \\ \varepsilon_{4tsk} \end{bmatrix}$$

$$\begin{bmatrix} \varepsilon_{1tsk} \\ \varepsilon_{2tsk} \\ \varepsilon_{3tsk} \\ \varepsilon_{4tsk} \end{bmatrix} \sim \mathcal{N}_4\left(0_4, VCV\right), \; V = \begin{bmatrix} \sigma_1 & 0 & 0 & 0 \\ 0 & \sigma_2 & 0 & 0 \\ 0 & 0 & \sigma_3 & 0 \\ 0 & 0 & 0 & \sigma_4 \end{bmatrix}, \; C = I_4$$

Ecriture matricielle du modèle M_3

On peut proposer le modèle M_3 :

$$\begin{bmatrix} F_{1tsk} \\ F_{2tsk} \\ F_{3tsk} \\ F_{4tsk} \end{bmatrix} = \mu \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} + \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \alpha_4 \end{bmatrix} + \beta_t \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} + \begin{bmatrix} \gamma_{1t} \\ \gamma_{2t} \\ \gamma_{3t} \\ \gamma_{4t} \end{bmatrix} + \xi_s \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} + \begin{bmatrix} \xi_{1s} \\ \xi_{2s} \\ \xi_{3s} \\ \xi_{4s} \end{bmatrix} + \xi_{ts} \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} + \begin{bmatrix} \varepsilon_{1tsk} \\ \varepsilon_{2tsk} \\ \varepsilon_{3tsk} \\ \varepsilon_{4tsk} \end{bmatrix}$$

$$\begin{bmatrix} \varepsilon_{1tsk} \\ \varepsilon_{2tsk} \\ \varepsilon_{3tsk} \\ \varepsilon_{4tsk} \end{bmatrix} \sim \mathcal{N}_4 \left(0_4, VCV\right),$$

$$V = \begin{bmatrix} \sigma_1 & 0 & 0 & 0 \\ 0 & \sigma_2 & 0 & 0 \\ 0 & 0 & \sigma_3 & 0 \\ 0 & 0 & 0 & \sigma_4 \end{bmatrix}, C = \begin{bmatrix} 1 & \sigma_{1,2} & \sigma_{1,3} & \sigma_{1,4} \\ \sigma_{1,2} & 1 & \sigma_{2,3} & \sigma_{2,4} \\ \sigma_{1,3} & \sigma_{2,3} & 1 & \sigma_{3,4} \\ \sigma_{1,4} & \sigma_{2,4} & \sigma_{3,4} & 1 \end{bmatrix}$$

Validation du modèle M_3 ?

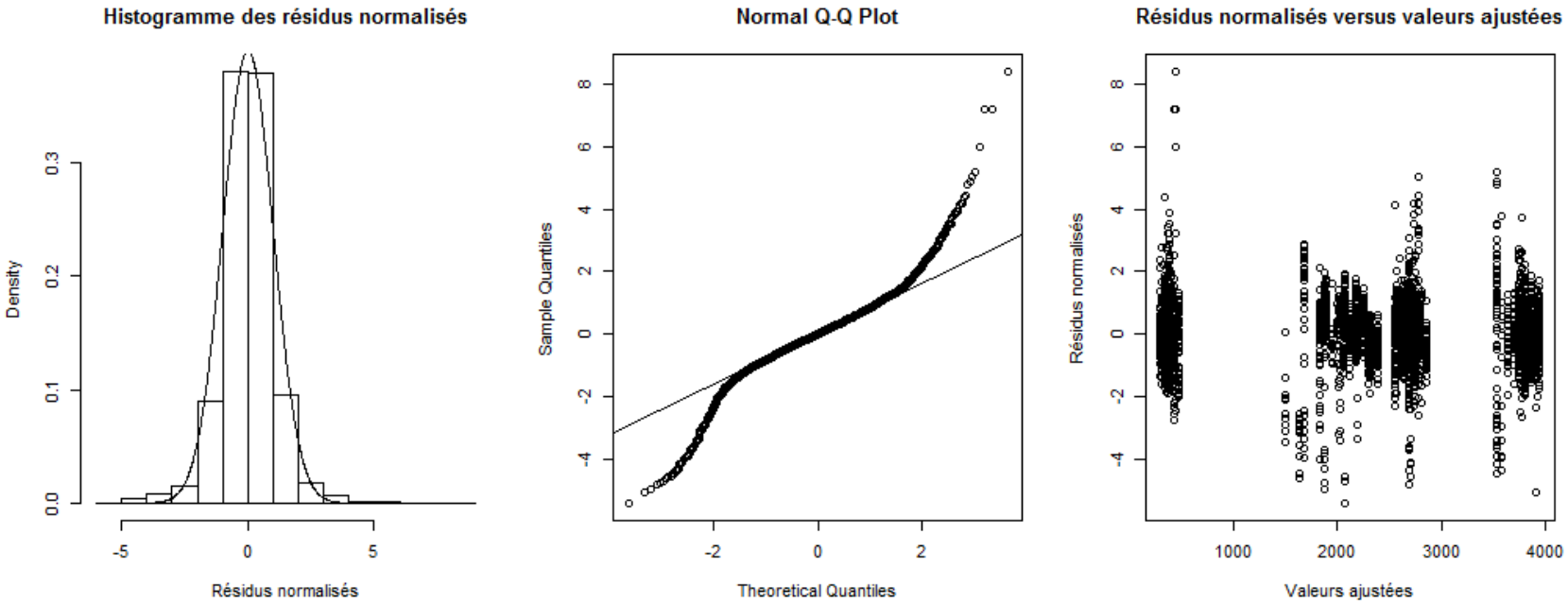


Figure: Graphiques diagnostiques pour le modèle M_3

- L'histogramme semble convenir.
- Le QQ-plot montre de grandes déviations sur les valeurs extrêmes.
- Le dernier graphique ne montre pas d'effets, mais de nombreuses valeurs aberrantes.

Validation du modèle M_3 ?

Pistes d'améliorations :

- Examiner les nombreuses valeurs aberrantes : sont-ce des données particulières, sont-elles fiables ?
- Chercher un facteur explicatif manquant ...

A titre pédagogique, on présente malgré tout la suite de la méthodologie à partir du modèle M_3 ...

Interprétation des estimateurs des coefficients

```
Fixed effects: F.mes ~ formant * Test
                  Value Std.Error   DF  t-value p-value
(Intercept)     371.174  41.39052 3711  8.96760  0.0000
formantF2      1555.196  58.43281 3711 26.61512  0.0000
formantF3      2330.465  56.76796 3711 41.05247  0.0000
formantF4      3443.460  57.68883 3711 59.69025  0.0000
Test2             3.912  15.79115 3711  0.24775  0.8043
formantF2:Test2 181.174  19.09356 3711  9.48876  0.0000
formantF3:Test2 -11.278   8.76430 3711 -1.28685  0.1982
formantF4:Test2 -45.840  15.32094 3711 -2.99198  0.0028
```

- La valeur moyenne des formants F_1 avant formation est estimée à 371.174.
- Avant formation, il existe des différences significatives entre les niveaux moyens de F_1 et F_2, F_1 et F_3, F_1 et F_4.
- Pour le formant F_1, il n'y a pas de différence significative entre avant et après formation.
- On ne peut rien dire de plus ...

Les valeurs cibles ont-elles été atteintes avant formation ?

Les valeurs cibles sont tirées de Georgetown et al. (2012)[5].

$$\begin{bmatrix} F_{1tsk} \\ F_{2tsk} \\ F_{3tsk} \\ F_{4tsk} \end{bmatrix} = \mu \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} + \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \alpha_4 \end{bmatrix} + \beta_t \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} + \begin{bmatrix} \gamma_{1t} \\ \gamma_{2t} \\ \gamma_{3t} \\ \gamma_{4t} \end{bmatrix} + \xi_s \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} + \begin{bmatrix} \xi_{1s} \\ \xi_{2s} \\ \xi_{3s} \\ \xi_{4s} \end{bmatrix} + \xi_{ts} \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} + \begin{bmatrix} \varepsilon_{1tsk} \\ \varepsilon_{2tsk} \\ \varepsilon_{3tsk} \\ \varepsilon_{4tsk} \end{bmatrix}$$

Pour répondre à cette question, on souhaite tester :

$$\begin{aligned} H_0 : \mu = 276 \quad &/ \quad H_1 : \mu \neq 276 \\ H_0 : \mu + \alpha_2 = 2091 \quad &/ \quad H_1 : \mu + \alpha_2 \neq 2091 \\ H_0 : \mu + \alpha_3 = 2579 \quad &/ \quad H_1 : \mu + \alpha_3 \neq 2579 \\ H_0 : \mu + \alpha_4 = 3826 \quad &/ \quad H_1 : \mu + \alpha_4 \neq 3826 \end{aligned}$$

Pour faire ces tests, on réalise des tests de contrastes (tests sur des combinaisons linéaires des coefficients).

[5] Georgeton, L., Paillereau, N., Landron, S., Gao, J., Kamiyama, T. (2012). Analyse formantique des voyelles orales du français en contexte isolé : à la recherche d'une référence pour les apprenants de FLE. In *L'apport de la voix chantée en intégration phonétique d'une langue étrangère* Besacier, L., Lecouteux, B. et Sérasset, G. (eds.). *Proceedings of the Joint Conference JEP-TALN-RECITAL, 1. XXIXièmes Journées d'Étude de la Parole. Grenoble ATALA/AFCP* : 145–152.

Les valeurs cibles ont-elles été atteintes après formation ?

$$\begin{bmatrix} F_{1tsk} \\ F_{2tsk} \\ F_{3tsk} \\ F_{4tsk} \end{bmatrix} = \mu \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} + \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \alpha_4 \end{bmatrix} + \beta_t \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} + \begin{bmatrix} \gamma_{1t} \\ \gamma_{2t} \\ \gamma_{3t} \\ \gamma_{4t} \end{bmatrix} + \xi_s \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} + \begin{bmatrix} \xi_{1s} \\ \xi_{2s} \\ \xi_{3s} \\ \xi_{4s} \end{bmatrix} + \xi_{ts} \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} + \begin{bmatrix} \varepsilon_{1tsk} \\ \varepsilon_{2tsk} \\ \varepsilon_{3tsk} \\ \varepsilon_{4tsk} \end{bmatrix}$$

On souhaite tester :

$$H_0 : \mu + \beta_{Ap} = 276 \quad / \quad H_1 : \mu + \beta_{Ap} \neq 276$$

$$H_0 : \mu + \alpha_2 + \beta_{Ap} + \gamma_{2,Ap} = 2091 \quad / \quad H_1 : \mu + \alpha_2 + \beta_{Ap} + \gamma_{2,Ap} \neq 2091$$

$$H_0 : \mu + \alpha_3 + \beta_{Ap} + \gamma_{3,Ap} = 2579 \quad / \quad H_1 : \mu + \alpha_3 + \beta_{Ap} + \gamma_{3,Ap} \neq 2579$$

$$H_0 : \mu + \alpha_4 + \beta_{Ap} + \gamma_{4,Ap} = 3826 \quad / \quad H_1 : \mu + \alpha_4 + \beta_{Ap} + \gamma_{4,Ap} \neq 3826$$

Pour faire ces tests, on réalise des tests de contrastes (tests sur des combinaisons linéaires des coefficients).

Différences avant/après formation ?

$$\begin{bmatrix} F_{1tsk} \\ F_{2tsk} \\ F_{3tsk} \\ F_{4tsk} \end{bmatrix} = \mu \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} + \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \alpha_4 \end{bmatrix} + \beta_t \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} + \begin{bmatrix} \gamma_{1t} \\ \gamma_{2t} \\ \gamma_{3t} \\ \gamma_{4t} \end{bmatrix} + \xi_s \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} + \begin{bmatrix} \xi_{1s} \\ \xi_{2s} \\ \xi_{3s} \\ \xi_{4s} \end{bmatrix} + \xi_{ts} \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} + \begin{bmatrix} \varepsilon_{1tsk} \\ \varepsilon_{2tsk} \\ \varepsilon_{3tsk} \\ \varepsilon_{4tsk} \end{bmatrix}$$

Les niveaux moyens avant et après formation sont :

$$\begin{aligned} F_1 : \mu \quad &/ \quad \mu + \beta_{Ap} \\ F_2 : \mu + \alpha_2 \quad &/ \quad \mu + \alpha_2 + \beta_{Ap} + \gamma_{2,Ap} \\ F_3 : \mu + \alpha_3 \quad &/ \quad \mu + \alpha_3 + \beta_{Ap} + \gamma_{3,Ap} \\ F_4 : \mu + \alpha_4 \quad &/ \quad \mu + \alpha_4 + \beta_{Ap} + \gamma_{4,Ap} \end{aligned}$$

Pour tester les différences entre avant et après formation, on souhaite donc tester :

$$\begin{aligned} H_0 : \beta_{Ap} = 0 \quad &/ \quad H_1 : \beta_{Ap} \neq 0 \\ H_0 : \beta_{Ap} + \gamma_{2,Ap} = 0 \quad &/ \quad H_1 : \beta_{Ap} + \gamma_{2,Ap} \neq 0 \\ H_0 : \beta_{Ap} + \gamma_{3,Ap} = 0 \quad &/ \quad H_1 : \beta_{Ap} + \gamma_{3,Ap} \neq 0 \\ H_0 : \beta_{Ap} + \gamma_{4,Ap} = 0 \quad &/ \quad H_1 : \beta_{Ap} + \gamma_{4,Ap} \neq 0 \end{aligned}$$

Les différences entre formants sont-elles similaires avant et après formation ?

Les niveaux moyens des différences entre formants, avant et après formation sont :

$$\begin{aligned} F_2 - F_1 : \alpha_2 \quad &/ \quad \alpha_2 + \gamma_{2,Ap} \\ F_3 - F_2 : \alpha_3 - \alpha_2 \quad &/ \quad \alpha_3 - \alpha_2 + \gamma_{3,Ap} - \gamma_{2,Ap} \\ F_4 - F_3 : \alpha_4 - \alpha_3 \quad &/ \quad \alpha_4 - \alpha_3 + \gamma_{4,Ap} - \gamma_{3,Ap} \end{aligned}$$

Pour tester les différences entre avant et après formation, on souhaite donc tester :

$$\begin{aligned} H_0 : \gamma_{2,Ap} = 0 \quad &/ \quad H_1 : \gamma_{2,Ap} \neq 0 \\ H_0 : \gamma_{3,Ap} - \gamma_{2,Ap} = 0 \quad &/ \quad H_1 : \gamma_{3,Ap} - \gamma_{2,Ap} \neq 0 \\ H_0 : \gamma_{4,Ap} - \gamma_{3,Ap} = 0 \quad &/ \quad H_1 : \gamma_{4,Ap} - \gamma_{3,Ap} \neq 0 \end{aligned}$$

Modèles linéaires généralisés

Modèles linéaires généralisés

Classe de modèles permettant de modéliser des variables d'autres natures

Exemples :

- Modèle de régression logistique : la variable réponse Y_i est une variable binaire (codée 0/1) et suit une loi de Bernoulli ;
- Modèle de régression Poissonnienne : la variable réponse Y_i est une variable de comptage (nombre entier) et suit une loi de Poisson ;
- ...

Jeu de données "Patients atteints de la maladie d'Alzheimer"

Ce jeu de données est issu des travaux de thèse de Diane Caussade[6]

- 10 patientes de la maladie d'Alzheimer et 10 sujets contrôles
- répètent des phrases, avec ou sans geste.
- On compte le nombre de pauses.

5 variables statistiques :

- Nombre de pauses : variable d'intérêt (ou réponse), discrète
- âge : variable explicative continue
- stade de la maladie (MMSE) : variable explicative qualitative ordonnée
- tâche (t = avec ou sans geste) : variable explicative qualitative nominale
- sujet (s) : variable qualitative nominale
- phrase (k) : variable qualitative nominale

[6] [Caussade] D. Caussade, N. Vallée, N. H. Bernardoni, J.-M. Colletta, S. Gerber, F. Letué, M.-J. Martinez (2016) Disfluences dans le vieillissement "normal" et la maladie d'Alzheimer : indices segmentaux, suprasegmentaux et gestuels. *Actes de la Conférence JEP*, 1, 182-190.

Modèle de régression logistique

On cherche à expliquer la variable binaire Y_{stk} définie par

$$\begin{aligned} Y_{stk} &= 1 \text{ si il y a au moins une pause dans la phrase} \\ &= 0 \text{ sinon.} \end{aligned} \quad (1)$$

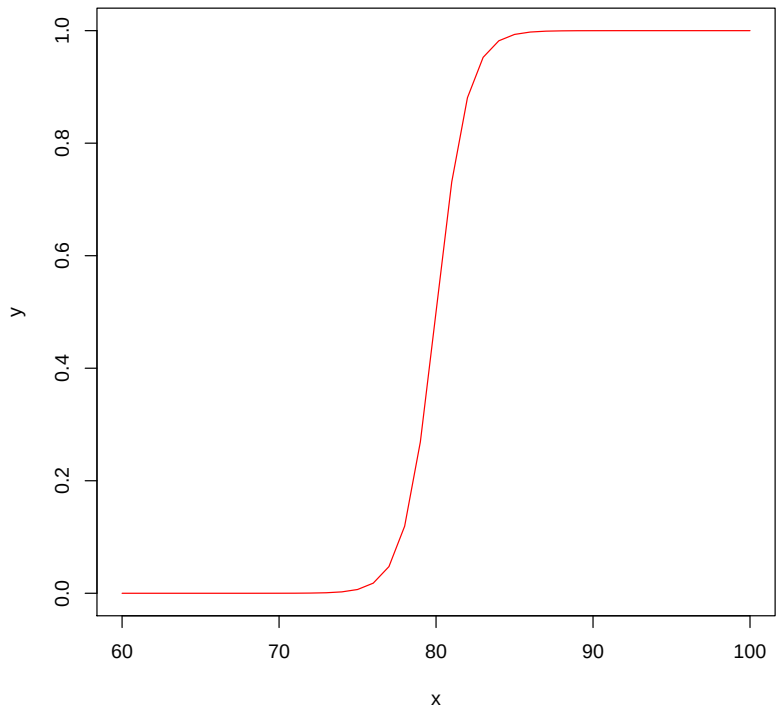
Dans un premier temps, on n'introduit que l'effet âge (x_s) dans le modèle.

Modèle de régression logistique

Le modèle de régression logistique est défini par

$$P(Y_{stk} = 1) = \frac{\exp(\beta_0 + \beta_1 x_s)}{1 + \exp(\beta_0 + \beta_1 x_s)} = logit^{-1}(\beta_0 + \beta_1 x_s).$$

Modèle mixte de régression logistique



Modèle mixte de régression logistique

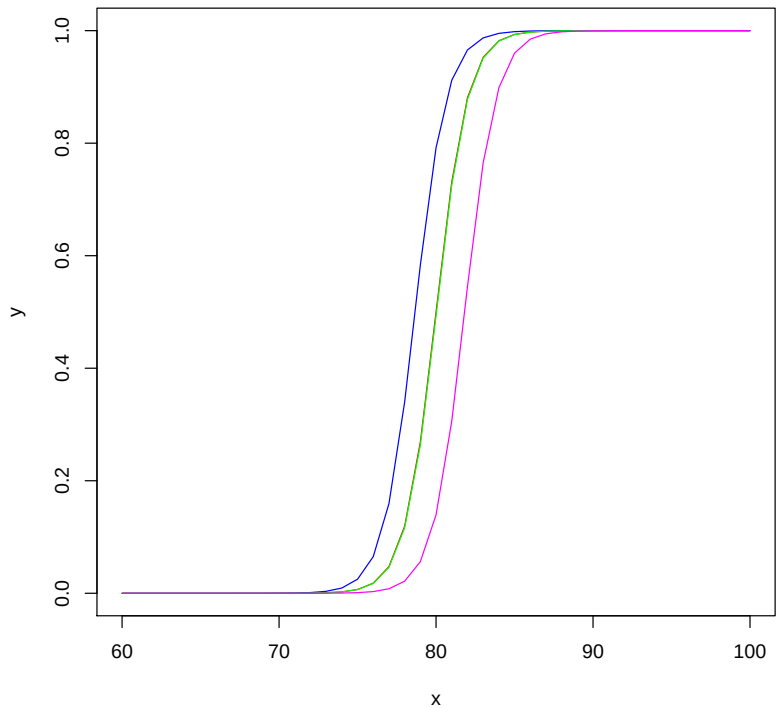
Pour tenir compte des effets individuels et des effets phrases, on ajoute deux effets aléatoires : $\xi_s \sim \mathcal{N}(0, \tau_1^2)$ et $\xi_k \sim \mathcal{N}(0, \tau_2^2)$.

Modèle mixte de régression logistique

Le modèle mixte de régression logistique est défini par

$$P(Y_{stk} = 1|\xi_s, \xi_k) = \frac{\exp(\beta_0 + \beta_1 x_s + \xi_s + \xi_k)}{1 + \exp(\beta_0 + \beta_1 x_s + \xi_s + \xi_k)} = logit^{-1}(\beta_0 + \beta_1 x_s + \xi_s + \xi_k).$$

Modèle mixte de régression logistique



Un exemple plus réaliste

$$P(Y_{stk} = 1|\xi_s, \xi_k) = logit^{-1}(\beta_0 + \beta_1 age_s + \beta_{2,t} + \beta_{3,mmse_s} + \gamma_{t,mmse_s} + \xi_s + \xi_k).$$

Fixed effects:

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	0.31453	1.27203	0.247	0.80470	
age	-0.02583	0.01528	-1.690	0.09105	.
taskvpsg	-0.73754	0.42425	-1.738	0.08213	.
mmse_fpat_leg	1.10132	0.35612	3.093	0.00198	**
mmse_fpat_mod	0.50768	0.36984	1.373	0.16984	
mmse_fpat_sev	0.85476	0.51336	1.665	0.09590	.
taskvpsg:mmse_fpat_leg	-0.48115	0.50836	-0.946	0.34391	
taskvpsg:mmse_fpat_mod	0.44090	0.50750	0.869	0.38498	
taskvpsg:mmse_fpat_sev	1.02117	0.61789	1.653	0.09840	.

Beaucoup de variables ne sont pas significatives. On effectue donc une sélection de variables.

Un exemple plus réaliste : sélection de variables

On commence par supprimer l'interaction MMSE-tâche.

$$P(Y_{stk} = 1|\xi_s, \xi_k) = logit^{-1}(\beta_0 \quad + \quad \beta_1 age_s + \beta_{2,t} + \beta_{3,mmse_s} + \xi_s + \xi_k).$$

```
Fixed effects:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)      0.16955    1.25437   0.135  0.89248
age             -0.02460    0.01514  -1.625  0.10420
taskvpsg        -0.61292    0.24816  -2.470  0.01352 *
mmse_fpat_leg    0.90905    0.30244   3.006  0.00265 **
mmse_fpat_mod    0.68504    0.30857   2.220  0.02641 *
mmse_fpat_sev    1.40665    0.39807   3.534  0.00041 ***
```

Un test de comparaison de modèle confirme que ce deuxième modèle est meilleur que le premier. Mais l'âge n'est toujours pas significatif.

Un exemple plus réaliste : sélection de variables

On supprime l'âge du modèle.

$$P(Y_{stk} = 1|\xi_s, \xi_k) = logit^{-1}(\beta_0 \quad + \quad \beta_{2,t} + \beta_{3,mmse_s} + \xi_s + \xi_k).$$

```
Fixed effects:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)    -1.8450     0.2694  -6.848 7.51e-12 ***
taskvpsg       -0.6158     0.2484  -2.479  0.01319 *
mmse_fpat_leg   0.9694     0.3140   3.087  0.00202 **
mmse_fpat_mod   0.6062     0.3161   1.918  0.05513 .
mmse_fpat_sev   1.2033     0.3937   3.056  0.00224 **
```

Un test de comparaison de modèle confirme que ce troisième modèle est meilleur que le deuxième. Maintenant, toutes les variables sont significatives : on s'arrête là.

Prédiction

Une fois les coefficients et les effets aléatoires du modèle ajustés, on peut calculer une probabilité prédite :

$$\hat{p}_{stk} = logit^{-1}(\hat{\beta}_0 \quad + \quad \hat{\beta}_{2,t} + \hat{\beta}_{3,mmse_s} + \hat{\xi}_s + \hat{\xi}_k).$$

Pour un seuil donné z, on calcule une prédiction $\hat{Y}_{stk}$ par :

$$\begin{aligned} \hat{Y}_{stk} &= 1 \text{ si } \hat{p}_{stk} > z \\ &= 0 \text{ sinon.} \end{aligned}$$

Sensibilité, spécificité

Sensibilité, spécificité

On appelle sensibilité le taux de vrais positifs : $\frac{card\{Y_{stk}=1 \cap \hat{Y}_{stk}=1\}}{card\{\hat{Y}_{stk}=1\}}$.

On appelle spécificité le taux de vrais négatifs : $\frac{card\{Y_{stk}=0 \cap \hat{Y}_{stk}=0\}}{card\{\hat{Y}_{stk}=0\}}$.

On appelle anti-spécificité le taux de faux négatifs : $\frac{card\{Y_{stk}=1 \cap \hat{Y}_{stk}=0\}}{card\{\hat{Y}_{stk}=0\}}$.

On construit la courbe ROC en traçant pour chaque seuil $z \in [0,1]$, la sensibilité en fonction de l'anti-spécificité. On calcule l'aire sous cette courbe (AUC).
Plus la courbe s'approche du coin supérieur gauche, plus l'aire est grande et meilleur est le modèle.

Courbe ROC

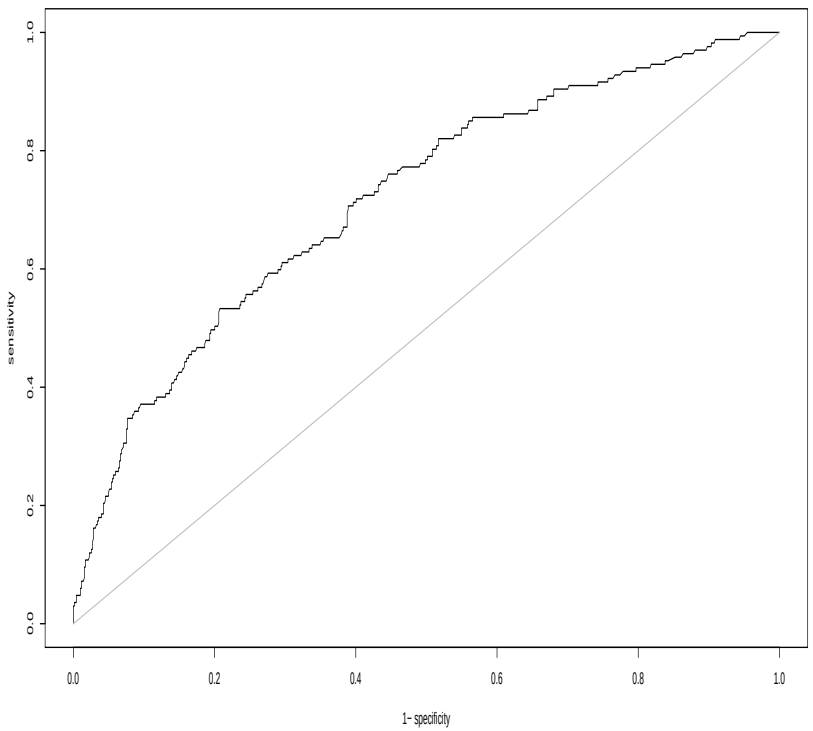


Figure: Courbe ROC : Sensibilité en fonction de l'anti-spécificité

L'aire sous la courbe est de 0.7181248.

Modèle de régression Poissonienne

On cherche à expliquer la variable de comptage N_{stk} qui représente le nombre de pauses dans la répétition de la phrase.

Loi de Poisson

On dit que N suit une loi de Poisson de paramètre λ si

$$P(N = n) = e^{-\lambda}\frac{\lambda^n}{n!}, n = 0, 1, \cdots$$

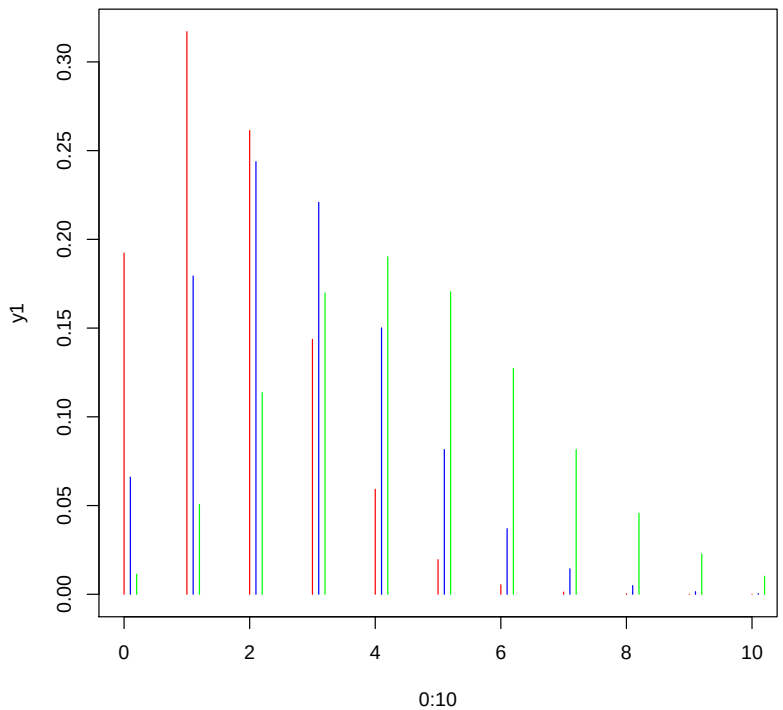
Dans un premier temps, on n'introduit que l'effet âge (x_s) dans le modèle.

Modèle de régression Poissonienne

Dans le modèle de régression Poissonienne, N_{stk} suit une loi de Poisson de paramètre

$$\lambda_s = e^{\beta_0+\beta_1 x_s}.$$

Modèle mixte de régression Poissonienne



Modèle mixte de régression Poissonienne

Pour tenir compte des effets individuels et des effets phrases, on ajoute deux effets aléatoires : $\xi_s \sim \mathcal{N}(0, \tau_1^2)$ et $\xi_k \sim \mathcal{N}(0, \tau_2^2)$.

Modèle mixte de régression Poissonienne

Dans le modèle mixte de régression Poissonienne, N_{stk} sachant le sujet et la phrase suit une loi de Poisson de paramètre

$$\lambda_{sk|\xi_s,\xi_k} = e^{\beta_0+\beta_1 x_s+\xi_s+\xi_k}.$$

Un exemple plus réaliste

$$\lambda_{sk|\xi_s,\xi_k} = exp(\beta_0 + \beta_1 age_s + \beta_{2,t} + \beta_{3,mmse_s} + \gamma_{t,mmse_s} + \xi_s + \xi_k).$$

Fixed effects:

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	3.28448	1.08819	3.018	0.002542	**
age	-0.06369	0.01342	-4.745	2.08e-06	***
taskvpsg	-0.70766	0.37233	-1.901	0.057352	.
mmse_fpat_leg	0.69217	0.30315	2.283	0.022415	*
mmse_fpat_mod	0.26721	0.34176	0.782	0.434300	
mmse_fpat_sev	1.01103	0.43223	2.339	0.019329	*
taskvpsg:mmse_fpat_leg	-0.04885	0.39528	-0.124	0.901646	
taskvpsg:mmse_fpat_mod	1.40780	0.40500	3.476	0.000509	***
taskvpsg:mmse_fpat_sev	1.23415	0.46322	2.664	0.007715	**

Ici, toutes les variables sont significatives : pas de sélection de variables.

Validation du modèle mixte de régression Poissonienne

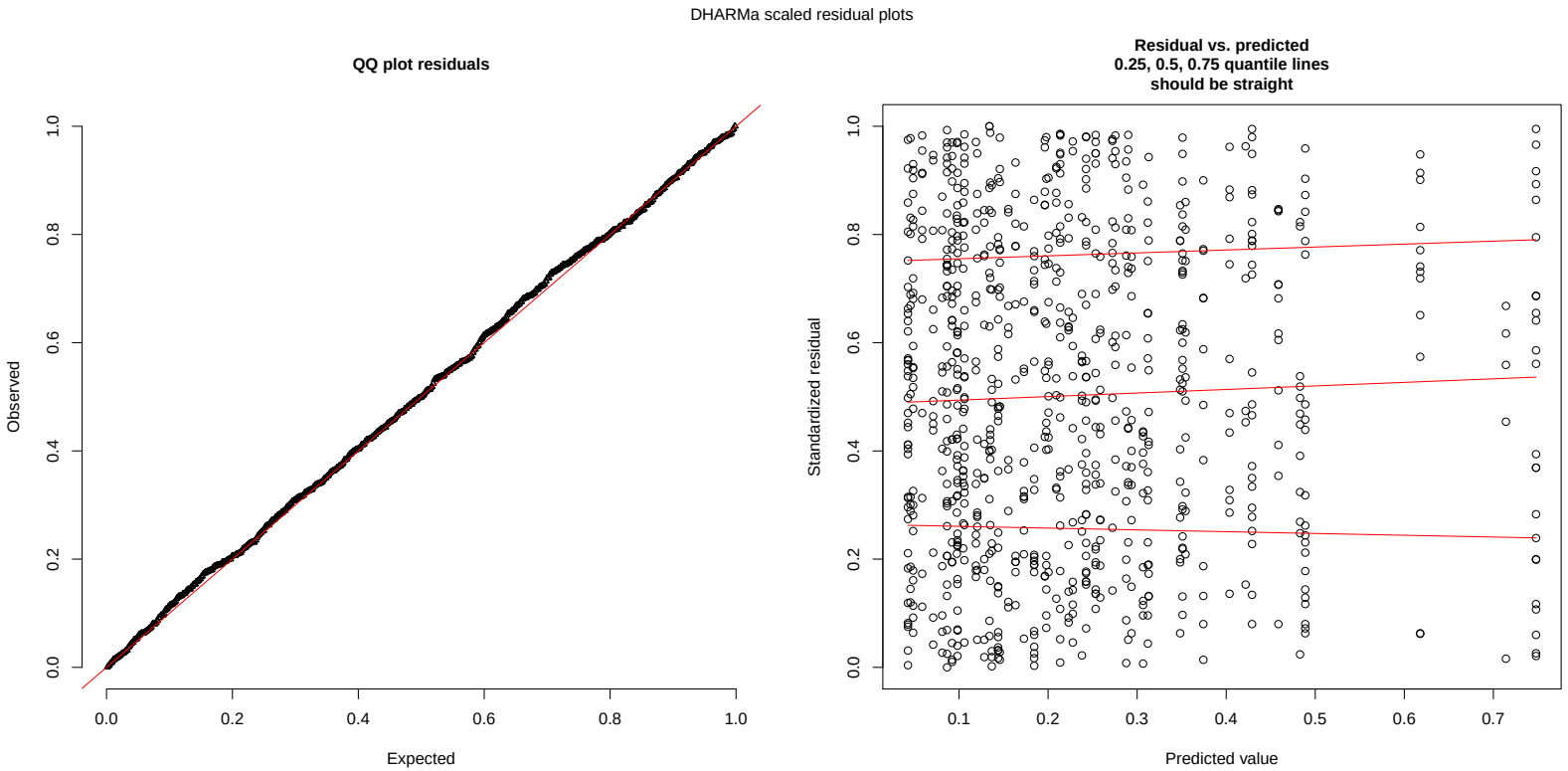


Figure: Graphiques diagnostiques pour le modèle mixte de régression Poissonienne.

Le modèle semble parfaitement ajusté (données simulées).

Validation du modèle mixte de régression Poissonienne

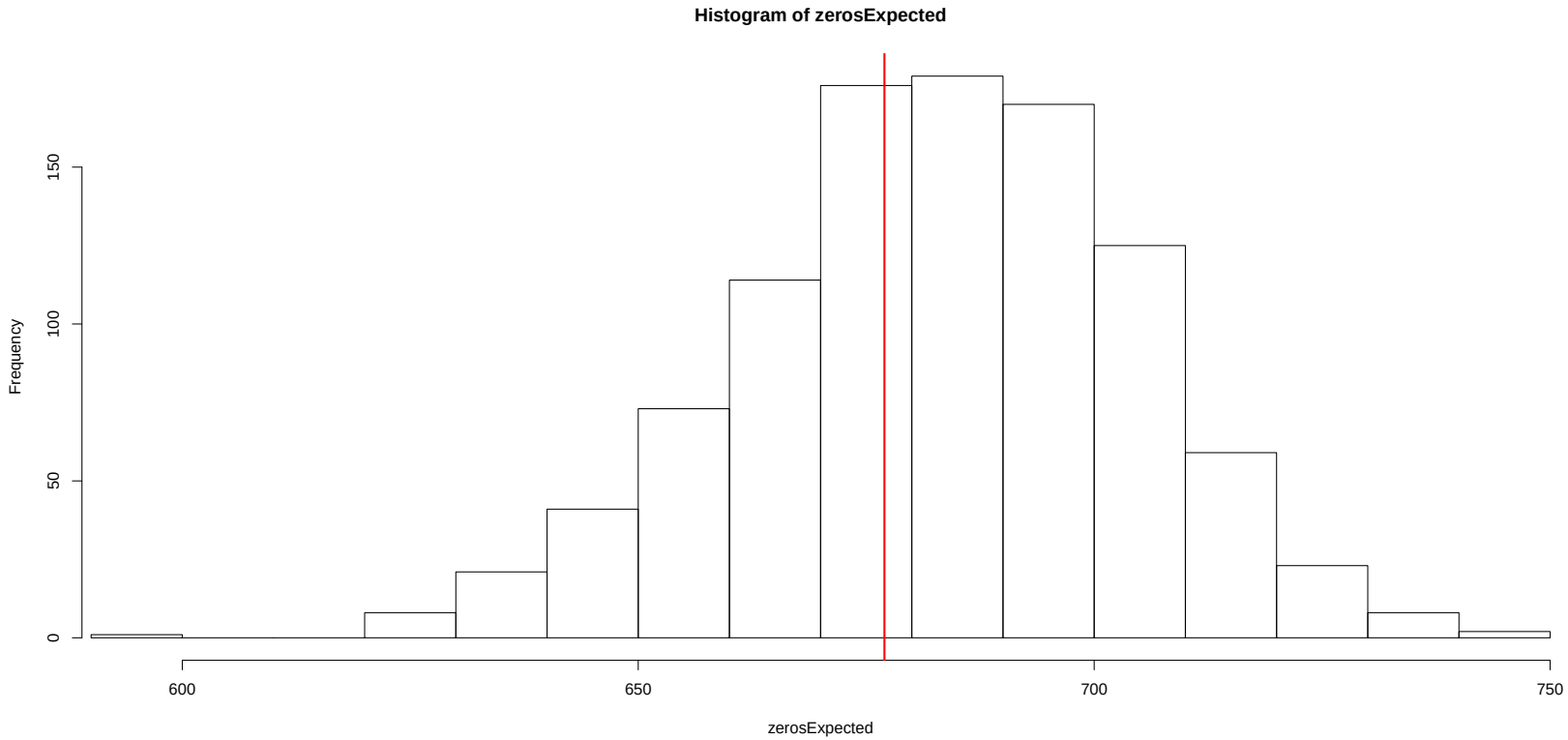


Figure: Graphique diagnostique pour le modèle mixte de régression Poissonienne.

Le modèle semble parfaitement ajusté (données simulées).

Jeu de données "Le cochon avec le chapeau rose"

Ce jeu de données est issu des travaux d'Anne Vilain et Coriandre Vilain [7]

- 21 sujets, enfants et adultes
- désignent (avec le doigt) et nomment l'intrus dans un ensemble d'objets
- On mesure la durée jusqu'à l'apex du geste.

4 variables statistiques :

- Durée du geste : variable d'intérêt (ou réponse), continue, positive
- groupe : enfants /adultes
- nombre de syllabes : 1, 2 ou 3
- sujet

[7] [Vilain] Vilain, C., Vilain, A., and Clarke, J. (2013). "the pig with the pink hat" : an experimental study on speech/gesture coordination during development. In *Proceedings of the TIGER conference*.

Modélisation par un modèle linéaire mixte

On commence par ajuster un modèle linéaire mixte M_0.

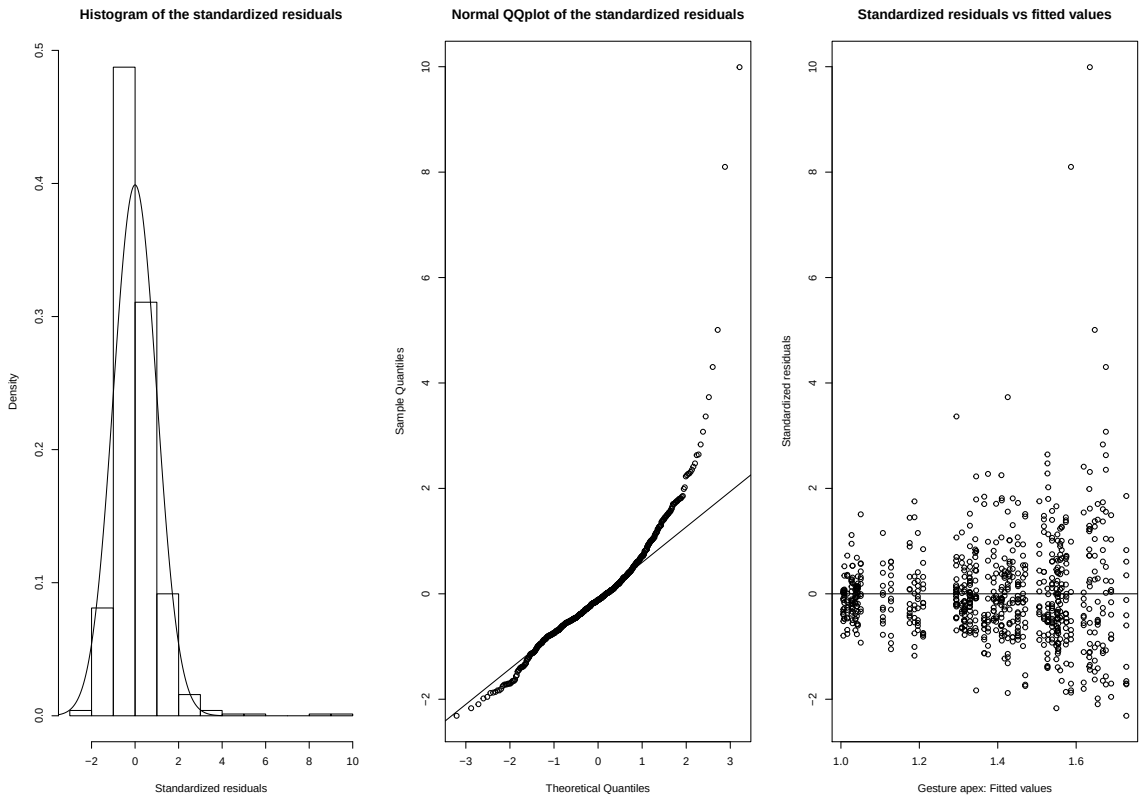


Figure: Graphiques diagnostiques pour le modèle M_0

Transformation de variable

On remarque que la distribution des résidus n'est pas gaussienne. C'est typique des variables de durées qui ont une queue de distribution lourde. Pour y remédier, on peut, dans un premier temps, considérer le log de la durée :

$$M_1 : \log(T_{s,nb,k}) = \mu + \alpha_g + \beta_{nb} + \gamma_{g,nb} + \xi_s + \varepsilon_{s,nb,k}.$$

Modélisation par un modèle log-linéaire mixte

Après sélection de variables, on obtient les résultats suivants :

```
Fixed effects: log(gesture_apex) ~ nb_syllables
                    Value  Std.Error  DF   t-value p-value
(Intercept)    0.29987642 0.03482566 730  8.610790  0.0000
nb_syllables2 -0.00505904 0.01282436 730 -0.394487  0.6933
nb_syllables3  0.02337550 0.01282436 730  1.822742  0.0688
```

L'effet d'un nombre de syllabes égal à 2 n'est pas significatif. Cela suggère de regrouper les modalités 1 et 2.

Modélisation par un modèle log-linéaire mixte

```
Fixed effects: log(gesture_apex) ~ (nb_syllables == 3)
                        Value  Std.Error  DF  t-value p-value
(Intercept)         0.29734690 0.03420708 731 8.692554  0.0000
nb_syllables == 3 0.02590502 0.01110001 731 2.333783  0.0199
```

Modélisation par un modèle log-linéaire mixte

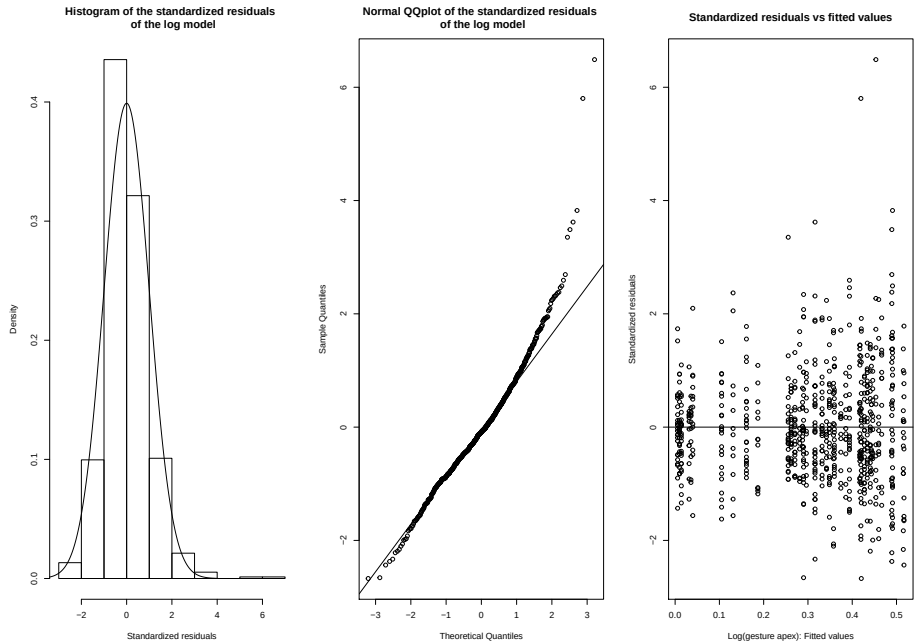


Figure: Graphiques diagnostiques pour le modèle M_1

Les graphiques sont améliorés, mais la distribution n'est toujours pas gaussienne et les résidus montrent une dispersion croissante avec la valeur ajustée. Le modèle n'est donc toujours pas adapté.

Outils probabilistes pour l'analyse des durées

Pour définir la loi des variables de durée, on utilise souvent le taux de risque définit par

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t | T \geq t)}{\Delta t}.$$

Il s'interprète comme la probabilité que l'événement survienne juste après t, sachant qu'il n'était pas survenu à la date t.

Modélisation par modèle de Cox mixte

Modèle de Cox

Le modèle de Cox suppose que le taux de risque individuel s'écrit

$$h_{g,nb}(t|\xi_s) = e^{\alpha_g + \beta_{nb} + \gamma_{g,nb} + \xi_s} h_0(t)$$

h_0 s'interprète comme le taux de risque d'un adulte moyen pour un mot d'une syllabe.

Modélisation par modèle de Cox mixte

	coef	se(coef)	se2	Chisq	DF	p
groupC	-0.36093	0.1322	0.1304	7.45	1	6.3e-03
nb_syllables2	0.05547	0.1141	0.1140	0.24	1	6.3e-01
nb_syllables3	-0.02160	0.1144	0.1143	0.04	1	8.5e-01
frailty.gaussian				16.68	1	4.4e-05
groupC:nb_syllables2	-0.17171	0.1850	0.1849	0.86	1	3.5e-01
groupC:nb_syllables3	-0.24377	0.1871	0.1870	1.70	1	1.9e-01

Certaines variables ne sont pas significatives : on procède à une sélection de variables.

Modélisation par modèle de Cox mixte

Modèle final :

	coef	se(coef)	se2	Chisq	DF	p
groupC	-0.501	0.0791	0.07631	40.12	1	2.4e-10
frailty.gaussian				16.44	1	5.0e-05

Il faut maintenant valider le modèle.

Validation du modèle de Cox mixte

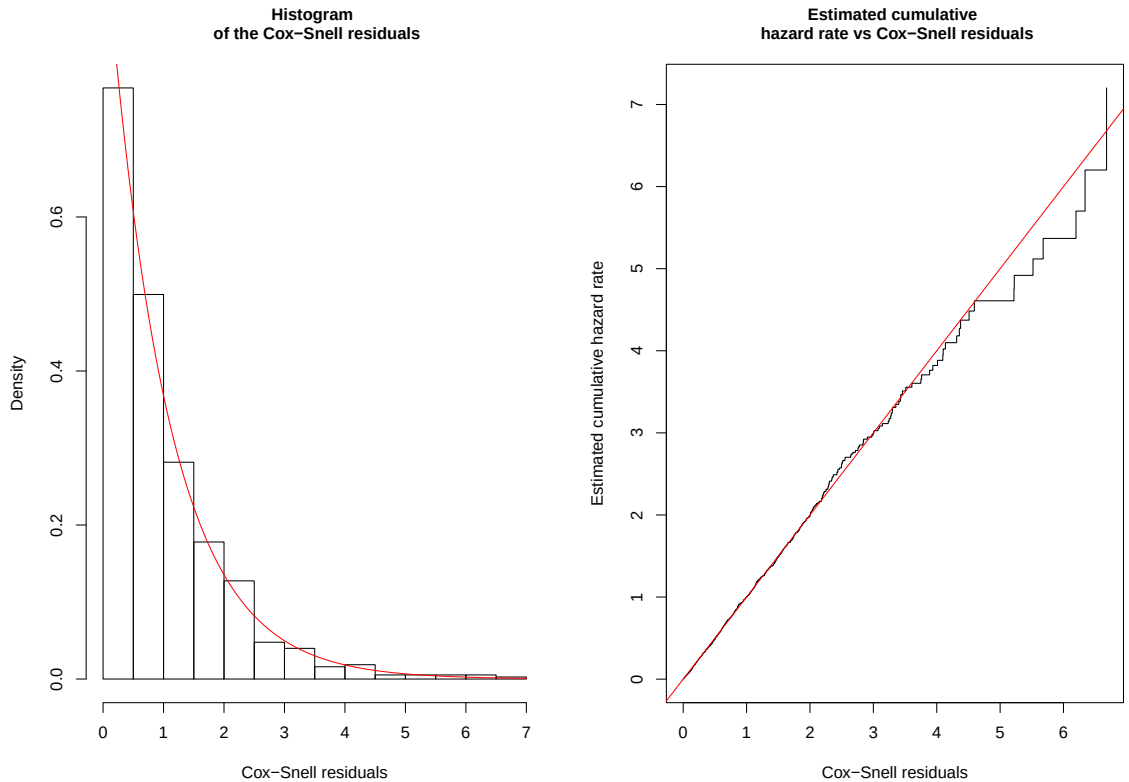


Figure: Graphiques diagnostiques pour le modèle de Cox mixte

Ce modèle semble bien adapté.

Conclusion

On a vu

- une grande variété de modèles statistiques
- adaptés à différents types de données (continues, discrètes, binaires,...)
- tenant compte des répétitions sur un même sujet
- qui permettent de tester des hypothèses variées.

Un message : toujours valider le modèle avant de tirer des conclusions des tests statistiques !


S. Cornaz Couffini (2014) L'apport de la voix chantée pour l'intégration phonético-phonologique d'une langue étrangère : application auprès d'italophones apprenants de FLE. *Thèse de l'Université Grenoble Alpes.*


Georgeton, L., Paillereau, N., Landron, S., Gao, J., Kamiyama, T. (2012). Analyse formantique des voyelles orales du français en contexte isolé : à la recherche d'une référence pour les apprenants de FLE. In *L'apport de la voix chantée en intégration phonétique d'une langue étrangère* Besacier, L., Lecouteux, B. et Sérasset, G. (eds.). *Proceedings of the Joint Conference JEP-TALN-RECITAL, 1. XXIXièmes Journées d'Étude de la Parole. Grenoble ATALA/AFCP* : 145–152.


D. Caussade, N. Vallée, N. H. Bernardoni, J.-M. Colletta, S. Gerber, F. Letué, M.-J. Martinez (2016) Disfluences dans le vieillissement "normal" et la maladie d'Alzheimer : indices segmentaux, suprasegmentaux et gestuels. *Actes de la Conférence JEP*, 1, 182-190.


Vilain, C., Vilain, A., and Clarke, J. (2013). "the pig with the pink hat" : an experimental study on speech/gesture coordination during development. In *Proceedings of the TIGER conference*.


Letué, F., Martinez, M.-J., Samson, A., Vilain, A., Vilain, C. (2017) Statistical methodology for the analysis of repeated duration data in speech, language and hearing research, *en révision*.