Université de Strasbourg
M2 - Statistique
Modèles linéaires généralisés

Année 2020/2021
Ségolen Geffray
geffray@math.unistra.fr

TD 2: modèles à effets mixtes

Exercice 1.
Dans chacun des cas suivants, déterminer le modèle marginal associé à chacun des modèles conditionnels suivants. Donner également leur forme matricielle.

1. $Y_{ij}|A_j \sim_{\text{indep}} \mathcal{N}(\mu + A_j, \sigma^2)$ pour $j = 1, ..., J$ et $i = 1, ..., n_j$ avec $A_j \sim_{iid} \mathcal{N}(0, \sigma_A^2)$.
2. $Y_{ij}|X_{ij}, A_j \sim_{\text{indep}} \mathcal{N}(\mu + \alpha X_{ij} + A_j, \sigma^2)$ pour $j = 1, ..., J$ et $i = 1, ..., n_j$ avec $A_j \sim_{iid} \mathcal{N}(0, \sigma_A^2)$.
3. $Y_{jk}|A_j \sim_{\text{indep}} \mathcal{N}(\mu + A_j + \beta_k, \sigma^2)$ pour $j = 1, ..., J$ et $k = 1, ..., K$ avec $A_j \sim_{iid} \mathcal{N}(0, \sigma_A^2)$ sous la contrainte $\beta_1 = 0$ ou de manière équivalente $\sum_{k=1}^{K} \beta_k = 0$.
4. $Y_{jk}|A_j, B_k \sim_{\text{indep}} \mathcal{N}(\mu + A_j + B_k, \sigma^2)$ pour $j = 1, ..., J$ et $k = 1, ..., K$ avec $A_j \sim_{iid} \mathcal{N}(0, \sigma_A^2)$, $B_k \sim_{iid} \mathcal{N}(0, \sigma_B^2)$ et avec l'hypothèse d'indépendance des A_j et des B_k.
5. $Y_{ijk}|A_j, C_{jk} \sim_{\text{indep}} \mathcal{N}(\mu + A_j + \beta_k + C_{jk}, \sigma^2)$ pour $j = 1, ..., J$, $k = 1, ..., K$ et $i = 1, ..., n_{jk}$ avec $A_j \sim_{iid} \mathcal{N}(0, \sigma_A^2)$, $C_{jk} \sim_{iid} \mathcal{N}(0, \sigma_{AB}^2)$ et avec l'hypothèse d'indépendance des A_j et des C_{jk} sous la contrainte $\beta_1 = 0$ ou de manière équivalente $\sum_{k=1}^{K} \beta_k = 0$.

Comment estimer les effets fixes? Comment prédire les effets aléatoires? Proposer un exemple concret dans lequel vous penseriez à utiliser chacun des modèles.

Exercice 2.
Afin d'évaluer l'efficacité d'une nouvelle molécule en prévention de l'hypertension artérielle, le dispositif expérimental suivant est élaboré. Sept fratries de souris sont sélectionnées, de tailles respectives 7, 8, 6, 6, 8, 7 et 8. La pression sanguine des souris est mesurée puis on leur administre une dose de la nouvelle molécule parmi 4 doses possibles (au prorata du poids de la souris), et, une heure après, leur pression sanguine est à nouveau mesurée. On supposera la linéarité des effets et la normalité de la loi conditionnelle de l'endogène.

1. Quel(s) modèle(s) envisageriez-vous d'ajuster sur les données collectées?
2. Comment traduire formellement l'hypothèse suivante: une hausse d'une unité de la dose de médicament a un effet sur la pression sanguine qui est le même quelle que soit la famille considérée?
3. Comment traduire formellement l'hypothèse suivante: l'effet du médicament sur la pression sanguine est non significatif?

Exercice 3.
Dans le contexte d'une étude à destination de l'industrie agro-alimentaire, on recherche les variables et facteurs influant le poids des veaux nouveaux-nés. On dispose pour cela de 100 vaches de 4 races différentes et de 5 taureaux de 2 races différentes. On envisage les possibles influences suivantes: race de la mère, race du père, âge de la mère, âge du père, sexe du veau. On supposera la linéarité des effets et la normalité de la loi conditionnelle de l'endogène. Comment procéderiez-vous?

Exercice 4.
Une laiterie industrielle possède plusieurs élevages de vaches laitières. Dans chacun des élevages, les vaches appartiennent aux deux races suivantes: béarnaise et jeirsiaise. L'industriel souhaite connaître les variables influençant la production laitière journalière. Pour chaque vache, il dispose des informations suivantes: production laitière journalière, race de la vache, élevage de la vache, alimentation (foin, ensilage de mais, granulé), poids de la vache, âge de la vache, nombre de vaches par unité de surface, nombre d'heures par jour passées à l'extérieur. Quel(s) modèle(s) proposeriez-vous dans le cadre de cette étude? Interpréter les paramètres introduits.

Exercice 5.
Dans le contexte d'une étude à destination de l'industrie agro-alimentaire, on recherche les variables et facteurs influant la croissance des poulets de Bresse (AOC). On dispose pour cela de 100 poulets élevés en partie en parcours extérieur. On mesure le poids du poulet à 1, 4, 8, 12 et 16 semaines. On envisage les possibles influences suivantes: régime alimentaire du poulet (4 types possibles), supplémentation en oligo-éléments, durée journalière d'exposition lumineuse, nombre de poulets par unité de surface. On supposera la linéarité des effets et la normalité de la loi conditionnelle de l'endogène. Comment procéderiez-vous?

Exercice 6.
Dans le contexte d'une étude sur la biodiversité du littoral, on réalise une étude sur le nombre d'espèces présentes dans l'estran. L'estran est la partie du littoral située entre les niveaux connus des plus hautes et des plus basses marées. Il constitue un biotope spécifique. Dix localités sont sélectionnées. Pour chacune d'entre elles, cinq sites de mesures de surfaces comparables sont échantillonnés. En chaque site, on relève le nombre d'espèces présentes, la hauteur du site de mesure par rapport au niveau moyen de marées, la pente du terrain, le plus haut niveau moyen de marée haute, le plus bas niveau moyen de marée basse, durée d'exondation (=retrait de la mer). Comment procéderiez-vous?

Exercice 7.
En foresterie, il est important de connaître les interactions entre les arbres et l'environnement. En particulier, les forestiers souhaitent évaluer l'impact de variables environnementales sur la propagation des maladies. Des arbres sont sélectionnés en différentes forêts du territoire français et sont examinés une fois par an pendant dix ans. Lors de chaque examen, le forestier détermine si l'arbre sélectionné est atteint par la maladie ou non. Sont également relevés: la pluviométrie cumulée lors de l'année écoulée (en mm), la régularité des précipitations lors de l'année écoulée (oui/non), l'ensoleillement quotidien moyen au cours de l'année écoulée (en h), la présence de températures extrêmes au cours de l'année écoulée (oui/non), le type de sol (calcaire, argileux, humifère, sableux), l'inclinaison du sol (nulle, faible, prononcée). Quel(s) modèle(s) proposeriez-vous dans le cadre de cette étude? Interpréter les paramètres introduits.

Exercice 8.
Dans le contexte de l'évaluation de l'impact de l'activité humaine sur l'environnement, on élabore le dispositif suivant. Un ruisseau est divisé en quatre ruisseaux expérimentaux. Le premier ruisseau est acidifié au H_2SO_4 et au HNO_3. Le deuxième ruisseau est acidifié seulement au H_2SO_4, le troisième ruisseau est acidifié seulement au HNO_3, le dernier ruisseau sert de témoin et n'a pas été acidifié. A 5 reprises, à chaque fois à une semaine d'intervalle, on mesure la quantité de chlorophylle (en mg/cm^3) dans chacun des trois ruisseaux.

		H_2SO_4	
		acidifié	non acidifié
HNO_3	acidifié	1.54	1.65
		1.50	1.50
		0.99	1.18
		1.52	2.19
		1.88	1.11
	non acidifié	1.55	3.28
		1.16	2.94
		1.99	3.50
		1.22	3.07
		1.95	3.04

Quelle est votre analyse? Ecrire l'équation et les hypothèses du modèle. Ecrire les tests de significativité des paramètres qui s'imposent.

Exercice 9.
Un médecin épidémiologiste analyse une étude européenne sur l'impact du tabagisme sur l'incidence du cancer du poumon. Les données ont été recueillies dans différents centres de soins. Soit Y_{ij} la variable binaire qui indique si la $i^{\text{ème}}$ personne du centre j est atteinte de cancer du poumon ou non. Soit $X_{ij}^{(1)}$ la variable qui indique si la $i^{\text{ème}}$ personne du centre j est non-fumeur, fumeur occasionnel, fumeur régulier ou gros fumeur, soit $X_{ij}^{(2)}$ la variable (quantitative) qui fournit la durée de tabagisme de la $i^{\text{ème}}$ personne du centre j et soit $X_{ij}^{(3)}$ la variable qui fournit le sexe de la $i^{\text{ème}}$ personne du centre j. Ecrire trois modèles de régression différents: l'un qui rend compte de l'influence de l'influence du centre de soins grâce à des effets fixes, l'autre qui rend compte de l'influence du centre de soins grâce à des effets aléatoires et le dernier qui néglige cette influence. Qu'en pensez-vous?

Exercice 10.
Supposons que l'on analyse les données d'un essai clinique pour comparer le nombre de crises d'épilepsie durant une période de 8 semaines avant le traitement (période durant laquelle le patient reçoit un placebo en une prise quotidienne) puis durant quatre périodes consécutives de 2 semaines (période durant laquelle le patient reçoit une nouvelle molécule en une prise quotidienne). On dispose des covariables suivantes: le traitement alloué au patient (placebo versus nouvelle molécule), la présence ou l'absence de maladie cérébrale décelable et la prédisposition génétique à la maladie (inspiré de Thall et Vail, 1990 ainsi que Diggle, Liang et Zeger, 1994). On souhaite évaluer l'effet du traitement. Quels modèles proposeriez-vous?

Exercice 11.
Un médecin épidémiologiste effectue une étude dans une région reculée des Andes. Il se rend

dans 12 villages de cette région, examine les enfants ayant entre 1 et 10 ans de chaque village et interroge leurs parents. Il cherche à expliquer le fait que les enfants soient vaccinés ou non. Il recueille pour cela les informations suivantes: âge de l'enfant, famille de l'enfant, nombre d'enfants dans la fratrie, place de l'enfant dans la fratrie, emploi occupé par la mère (0: aucun, 1: travail agricole, 2: travail manuel, artisanat, 3: autre), emploi occupé par le père (0: aucun, 1: travail agricole, 2: travail manuel, artisanat, 3: autre), nombre d'années de scolarisation de la mère, nombre d'années de scolarisation du père, distance au centre de santé le plus proche. Parmi les enfants observés, certains appartiennent à la même fratrie selon la répartition suivante:

nombre d'enfants observés d'une même fratrie	1	2	3	4	5	6
effectif	7	13	15	10	9	8

Quel(s) modèle(s) proposeriez-vous dans le cadre de cette étude? Comment l'interpréteriez-vous?

Exercice 12.
Considérons un modèle de régression gaussien avec le lien logarithme incluant une variable explicative quantitative $X^{(1)}$ ainsi qu'un facteur à effets mixtes, utilisant la loi gaussienne pour la partie effets aléatoires.

1. Ecrire l'équation et les hypothèses du modèle linéaire mixte conditionnel.
2. Déterminer le modèle marginal.
3. Comment estimer les effets fixes?
4. Comment prévoir les effets aléatoires?

Exercice 13.
Considérons un modèle de régression de Poisson avec le lien canonique incluant deux variables explicatives quantitatives $X^{(1)}$ et $X^{(2)}$, un facteur $X^{(3)}$ à 3 modalités ainsi qu'un facteur Z à effets aléatoires gaussiens.

1. Ecrire l'équation et les hypothèses du modèle linéaire mixte conditionnel.
2. Déterminer le modèle marginal.
3. Comment estimer les effets fixes?
4. Comment prévoir les effets aléatoires?

Exercice 14.
Considérons un modèle de régression exponentielle avec le lien logarithme incluant p variables explicatives quantitatives à effets fixes ainsi qu'un facteur Z à effets aléatoires gaussiens.

1. Ecrire l'équation et les hypothèses du modèle linéaire mixte conditionnel.
2. Déterminer le modèle marginal.
3. Comment estimer les effets fixes?
4. Comment prévoir les effets aléatoires?

Exercice 15.
Considérons un modèle de régression de Bernoulli avec le lien probit incluant p variables explicatives quantitatives à effets fixes ainsi qu'un facteur Z à effets aléatoires gaussiens.

1. Ecrire l'équation et les hypothèses du modèle linéaire mixte conditionnel.
2. Déterminer le modèle marginal.
3. Comment estimer les effets fixes?
4. Comment prévoir les effets aléatoires?

Exercice 16.
Déterminer le modèle marginal associé à chacun des modèles conditionnels suivants. Interpréter.

1. $Y_{ik} = \beta_0 + \beta_1 X_{ik} + A_i + B_i X_{ik} + \varepsilon_{ik}$ pour $k = 1, ..., K$ et $i = 1, ..., I$ avec $\begin{pmatrix} A_i \\ B_i \end{pmatrix} \sim_{iid} \mathcal{N}_2\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_A^2 & \rho\sigma_A\sigma_B \\ \rho\sigma_A\sigma_B & \sigma_B^2 \end{pmatrix}\right)$ et avec $\varepsilon_{ik} \sim_{iid} \mathcal{N}(0, \sigma^2)$ indépendants des (A_i, B_i).
2. Recommencer en supposant maintenant que $\begin{pmatrix} A_i \\ B_i \end{pmatrix} \sim_{iid} \mathcal{N}_2\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_A^2 & 0 \\ 0 & \sigma_B^2 \end{pmatrix}\right)$.
3. Recommencer mais avec une régression de Poisson.

Exercice 17.
Pour chacun des modèles conditionnels suivants, écrire le modèle marginal puis déterminer le meilleur prédicteur ainsi que le meilleur prédicteur linéaire des effets aléatoires à partir du vecteur réponse noté $\mathbb{Y} = (Y_{ij})_{\substack{1 \le j \le J \\ 1 \le i \le n_j}}$. On notera $\mathbb{X}_j = (1, X_j^{(1)}, ..., X_j^{(p)})$ un vecteur de prédicteurs quantitatifs (que l'on a précédés de la constante 1).

1. $Y_{ij}|\mu_j \sim_{indep} \mathcal{N}(\mu_j, \sigma^2)$ pour $j = 1, ..., J$ et $i = 1, ..., n_j$ avec $\mu_j|\mathbb{X}_j \sim_{indep} \mathcal{N}(\mathbb{X}_j.\beta, \sigma_A^2)$.
2. $Y_{ij}|\lambda_j \sim_{indep} \mathcal{P}(\lambda_j)$ pour $j = 1, ..., J$ et $i = 1, ..., n_j$ avec $\lambda_j|\mathbb{X}_j \sim_{indep} \Gamma(a_j, b_j)$ où $\log\left(\frac{a_j}{b_j}\right) = \mathbb{X}_j.\beta$ avec $a_j, b_j > 0$.

Exercice 18.
Considérons un modèle de régression de Poisson dit "zero-inflated", ce qui se définit comme un modèle de mélange entre une masse de Dirac en 0 et un modèle de régression de Poisson et ajoutons-y un effet aléatoire gaussien afin de modéliser une dépendance en cluster dans les données. Plus précisément, soit Y_{ij} la variable réponse de l'individu i du cluster j pour $i = 1, ..., I$ et $j = 1, ..., J$ et soit $\mathbb{X}_{i,j} = (1, X_{ij}^{(1)}, ..., X_{ij}^{(p)})$ où $X_{ij}^{(1)}, ..., X_{ij}^{(p)}$ sont des covariables à effets fixes pour l'individu i du cluster j. Soient A_j pour $j = 1, ..., J$ des variables aléatoires i.i.d. de loi $\mathcal{N}(0, \sigma^2)$. Posons un modèle conditionnel ainsi:

$$Y_{ij}|\mathbb{X}_{i,j}, A_j \sim \begin{cases} 0 & \text{avec probabilité } \pi_{ij} \\ \mathcal{P}(\lambda_{ij}) & \text{avec probabilité } (1 - \pi_{ij}) \end{cases}$$

avec $\log(\lambda_{ij}) = \mathbb{X}_{ij}.\beta + A_j$ et $\text{logit}(\pi_{ij}) = \mathbb{X}_{ij}.\gamma$. Supposons, de plus, que les $(Y_{ij}, X_{ij}^{(1)}, ..., X_{ij}^{(p)})$ sont indépendants conditionnellement aux A_j. Déterminer le modèle marginal. Comment interpréter ce modèle?

Exercice 19.
Une étude est réalisée sur l'activité journalière de la faune microbienne dans les mares. On sélectionne 20 mares suffisamment distantes pour que l'activité de la faune microbienne d'une mare n'ait aucune influence sur l'activité de la faune microbienne d'une autre. Pour chaque mare, quatre jours de suite, on mesure à chaque heure cette activité par la quantité de méthane dégagée (en cm^3). Notons $Y_{i,k}$ la quantité de méthane dégagée au cours de la $k^{\text{ème}}$ heure par la $i^{\text{ème}}$ mare.

1. Dans un premier temps, le statisticien envisage le modèle suivant:

$$Y_{i,k} = \beta_0 + \beta_1 \sin\left(\frac{2\pi k}{24}\right) + \varepsilon_{i,k}$$

en supposant que les vecteurs $(\varepsilon_{i,k})_{k=1,...,96}$ sont indépendants pour $i = 1, ..., 20$ et que chaque vecteur $(\varepsilon_{i,k})_{k=1,...,96}$ suit la loi $\mathcal{N}_{96}(0, \sigma^2 R(\rho))$ où $R(\rho) = \left[\rho^{|i-j|}\right]_{i,j=1,...,96}$ avec $0 < \rho < 1$ et $\sigma^2 > 0$. Comment interpréter ce modèle? Comment estimer β_0 et β_1?

2. Il se trouve que le statisticien dispose des covariables suivantes:

   - situation (à savoir: la mare est-elle sous couvert végétal arboré, ou entourée de végétation de type roseaux et grandes herbes, ou entourée d'un milieu ras)
   - volume de la mare (en m^3)
   - pH de l'eau (sur une échelle entre 0 et 14)
   - concentration en nitrates (en $\mu g/l$)
   - température (en degré Celsius)
   - lumière incidente (en lumen par m^2)

   Proposer un modèle à effets mixtes adapté à cette étude. Déterminer le modèle marginal associé.

3. Recommencer en supposant que l'étude porte à présent sur la biodiversité dans les mares. Pour chaque mare, on mesure cette fois le nombre d'espèces de micro-organismes par litre d'eau le premier jour de chaque mois et ce pendant un an. On notera ici $Y_{i,k}$ le nombre d'espèces de micro-organismes par litre relevé le premier jour du $k^{\text{ème}}$ mois pour la $i^{\text{ème}}$ mare.

Exercice 20.
Détailler le design d'une étude par simulation de la robustesse au choix de la loi des effets aléatoires dans un modèle linéaire gaussien à effets mixtes.

Exercice 21.
Ecrire les équations et hypothèses définissant les modèles suivants ajustés avec le package `lme4` du logiciel `R`.

1. `lmer(y~1+(1|F1)+(1|F2))`
2. `lmer(y~1+F1+(1+F1|F2))`
3. `lmer(y~1+F1+(1|F2)+(0+F1|F2))`

4. `glmer(y~F1*X2+I(X2^2)+F3+(1|F3:F4),family=binomial)`

Les variables dont le nom commence par 'F' sont des facteurs tandis que les variables dont le nom commence par 'X' sont des variables quantitatives.

Exercice 22.

Ci-dessous sont présentés quelques éléments d'analyse du jeu de données `Penicillin` disponible dans le logiciel `R` recueillies lors d'une étude sur la variabilité des échantillons de pénicilline: 144 observations sont disponibles. Les données contiennent une variable à expliquer:
`diameter` : diamètre de la zone d'inhibition des bactéries (exprimé en mm),
ainsi que les variables:
`plate`: identifiant du milieu nutritif mis dans la boite de Petri sur laquelle l'observation est effectuée, facteur de niveaux 'a' à 'x'
et
`sample`: identifiant de l'échantillon de pénicilline avec lequel l'observation est effectuée, facteur de niveaux 'A' à 'F'.

1. Comment interpréter le graphique exploratoire présenté en figure 1 obtenu grace aux instructions suivantes:

```
dotplot(reorder(plate, diameter) ~ diameter, Penicillin, groups = sample,
            ylab = "Plate", xlab = "Diameter of growth inhibition zone (mm)",
            type = c("p", "a"), auto.key = list(columns = 3, lines = TRUE,
            title = "Penicillin sample"),lwd=3)
```

2. Ecrire les équations et hypothèses définissant le modèle ajusté ci-dessous avec le logiciel `R` après avoir introduit les variables aléatoires nécessaires à la formalisation du problème.

```
> myfit <- lmer(diameter ~ 1+(1|plate) + (1|sample),data=Penicillin)
> summary(myfit)
Linear mixed model fit by REML
Formula: diameter ~ 1 + (1 | plate) + (1 | sample)
   Data: Penicillin
   AIC   BIC logLik deviance REMLdev
 338.9 350.7 -165.4    332.3   330.9
Random effects:
 Groups   Name        Variance Std.Dev.
 plate    (Intercept) 0.71691  0.84670
 sample   (Intercept) 3.73092  1.93156
 Residual             0.30242  0.54992
Number of obs: 144, groups: plate, 24; sample, 6

Fixed effects:
            Estimate Std. Error t value
(Intercept)  22.9722     0.8085   28.41
```

3. Déterminer le modèle marginal associé au modèle conditionnel posé en question 2.

Exercice 23.

Ci-dessous est présentée une étude par simulations de Monte-Carlo.

1. Ecrire les équations et hypothèses définissant le modèle étudié ci-dessous avec le logiciel R après avoir introduit les variables aléatoires nécessaires à la formalisation du problème.
2. Ecrire les hypothèses testées.
3. Qu'illustrent ces simulations?
4. Quelle(s) conclusion(s) en tirer?

```
> library(lme4)
> library(HLMdiag)
# contient la fonction varcomp.mer qui extrait les composantes de la variance

> M<-1000
>
> beta0_hat <- rep(NA,M)
> beta1_hat <- rep(NA,M)
> sigma2A_hat <- rep(NA,M)
> sigma2_hat <- rep(NA,M)
> sd_hat_beta0_hat <- rep(NA,M)
> sd_hat_beta1_hat <- rep(NA,M)
>
> for(m in 1:M)
+  {
+   x<- rnorm(n=300,mean=10,sd=2)
+   eps<- rnorm(n=300,mean=0,sd=1)
+   a<- rexp(n=30,rate=0.1)
+   y<-rep(NA,300)
+   Id<-rep(NA,300)
+   for (i in 1:30)
+      { for (j in 1:10)
+         { y[(i-1)*10+j]<- -10+x[(i-1)*10+j]+a[i]+eps[(i-1)*10+j]
+           Id[(i-1)*10+j]<- i
+           }
+        }
+
+   myfit <- lmer(y ~ x + (1|Id))
+   myout <- as.list(summary(myfit))
+
+   beta0_hat[m] <- myout$coef[,1][1]
+   beta1_hat[m] <- myout$coef[,1][2]
+   sigma2A_hat[m] <- varcomp.mer(myfit)[2]
+   sigma2_hat[m] <- varcomp.mer(myfit)[1]
+   sd_hat_beta0_hat[m] <- myout$coef[,2][1]
+   sd_hat_beta1_hat[m] <- myout$coef[,2][2]
+  }
>
> mean(beta0_hat)
[1] 0.000398016
> mean(beta1_hat)
[1] 0.9990194
```

```
> sd(beta0_hat)
[1] 1.815527
> sd(beta1_hat)
[1] 0.0307616
> mean(sigma2A_hat)
[1] 99.65518
> mean(sigma2_hat)
[1] 0.9985273
>
> aux0<- (beta0_hat - rep(-10,M))/sd_hat_beta0_hat
> aux1<- (beta1_hat - rep(-10,M))/sd_hat_beta1_hat
> print(shapiro.test(aux0))

        Shapiro-Wilk normality test

data:  aux0
W = 0.9902, p-value = 3.09e-06

> print(shapiro.test(aux1))

        Shapiro-Wilk normality test

data:  aux1
W = 0.9976, p-value = 0.153
```

Exercice 24.
Une étude est réalisée afin d'apprécier l'efficacité des messages de prévention sur les conséquences du tabac et de la drogue sur la santé en direction des adolescents. Chaque adolescent peut recevoir le message sous la forme d'une intervention effectuée dans le cadre scolaire et/ou sous la forme de campagne télévisée. Le groupe contrôle ne reçoit pas de message de prévention. Au total, 1600 collégiens dans 135 classes de 28 collèges sont inclus dans l'étude. A l'issue de l'étude, on mesure un score obtenu par l'adolescent à un test de connaissances sur les conséquences du tabac et de la drogue sur la santé. Les variables recueillies sont donc les suivantes:

- `TV`= 1 si l'adolescent reçoit le message sous la forme d'une campagne télévisée, = 0 dans le cas contraire,
- `SC`= 1 si l'adolescent reçoit le message sous la forme d'une intervention effectuée dans le cadre scolaire, = 0 dans le cas contraire,
- `PTHK`= score obtenu par l'adolescent à un test de connaissances sur les conséquences du tabac et de la drogue sur la santé avant l'entrée dans l'étude,
- `y`= score obtenu par l'adolescent à un test de connaissances sur les conséquences du tabac et de la drogue sur la santé à l'issue de l'étude.

Le logiciel `R` permet d'obtenir les résultats suivants.

```
>str(Smoking)
'data.frame':   1600 obs. of  6 variables:
```

```
 $ school: Factor w/ 28 levels "193","194","196",..: 9 9 9 9 9 9 9 9 9 9 ...
 $ class : Factor w/ 135 levels "193101","194101",..: 27 27 27 27 27 27 27 27 27 27 ...
 $ SC    : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
 $ TV    : Factor w/ 2 levels "0","1": 0 0 0 0 0 0 0 0 0 0 ...
 $ PTHK  : int  2 4 4 3 3 4 2 4 5 3 ...
 $ y     : int  3 4 3 4 4 3 2 4 5 4 ...
>
> myfit <- lmer(y ~ PTHK + SC + TV + school + class + (1|school) + (1|class),
+ data=Smoking)
fixed-effect model matrix is rank deficient so dropping 29 columns / coefficients
>
> myfit <- lmer(y ~ PTHK + SC + TV + school + (1|school) + (1|class),data=Smoking)
fixed-effect model matrix is rank deficient so dropping 2 columns / coefficients
Message d'avis :
In checkConv(attr(opt, "derivs"), opt$par, ctrl = control$checkConv,  :
  Model is nearly unidentifiable: large eigenvalue ratio
 - Rescale variables?
>
> myfit <- lmer(y ~ PTHK + SC + TV + (1|school) + (1|class),data=Smoking)
> summary(myfit)
Linear mixed model fit by REML ['lmerMod']
Formula: y ~ PTHK + SC + TV + (1 | school) + (1 | class)
   Data: Smoking

REML criterion at convergence: 5374.3

Scaled residuals:
    Min      1Q  Median      3Q     Max
-2.5202 -0.6975 -0.0177  0.6875  3.1630

Random effects:
 Groups   Name        Variance Std.Dev.
 class    (Intercept) 0.06853  0.2618
 school   (Intercept) 0.03925  0.1981
 Residual             1.60108  1.2653
Number of obs: 1600, groups:  class, 135; school, 28

Fixed effects:
            Estimate Std. Error t value
(Intercept)  1.78493    0.11295  15.803
PTHK         0.30524    0.02590  11.786
SC1          0.47147    0.11330   4.161
TV1          0.01956    0.11330   0.173

Correlation of Fixed Effects:
     (Intr) PTHK   SC1
PTHK -0.493
SC1  -0.503  0.025
TV1  -0.521  0.015 -0.002
```

1. Ecrire les équations et hypothèses définissant le modèle ajusté après avoir introduit les

variables aléatoires nécessaires à la formalisation du problème.

2. Donner les estimations réalisées. Interpréter les résultats produits.

3. Déterminer le modèle marginal.

4. Quelle alternative aurait-on pu proposer pour la partie "effets aléatoires"?

Exercice 25.

Ci-dessous sont présentés quelques éléments d'analyse du jeu de données `sleepstudy` disponible dans le logiciel `R` recueillies lors d'une étude sur la privation répétée de sommeil: 18 individus sont observés à 10 occasions. Les données contiennent une variable à expliquer:
`Reaction` : temps de réaction observé (exprimé en ms),
ainsi que les variables:
`Days`: nombre de jours écoulés depuis le début de la privation de sommeil,
et
`Subject`: identifiant de l'individu sur lequel l'observation est effectuée.

1. Comment interpréter les graphiques exploratoires présentés en figures 2 et 3 obtenus grace aux instructions suivantes:

```
xyplot(Reaction ~ Days | Subject, data=sleepstudy, type = c("g","p","r"),
       index = function(x,y) coef(lm(y ~ x))[1],
       xlab = "Days of sleep deprivation",
       ylab = "Average reaction time (ms)", aspect = "xy")
plot(confint(lmList(Reaction~Days|Subject,data=sleepstudy),pooled=TRUE),order=1)
```

2. Ecrire les équations et hypothèses définissant le modèle ajusté ci-dessous avec le logiciel `R` après avoir introduit les variables aléatoires nécessaires à la formalisation du problème.

```
> myfit <- lmer(Reaction ~ Days + (1|Subject) + (0+Days|Subject), sleepstudy)
> summary(myfit)

Linear mixed model fit by REML
Formula: Reaction ~ Days + (1 | Subject) + (0 + Days | Subject)
   Data: sleepstudy
  AIC  BIC logLik deviance REMLdev
 1754 1770 -871.8     1752    1744

Random effects:
 Groups   Name        Variance Std.Dev.
 Subject  (Intercept) 627.568  25.0513
 Subject  Days         35.858   5.9882
 Residual             653.584  25.5653
Number of obs: 180, groups: Subject, 18

Fixed effects:
            Estimate Std. Error t value
(Intercept)  251.405      6.885   36.51
Days          10.467      1.559    6.71
```

```
Correlation of Fixed Effects:
     (Intr)
Days -0.184
```

3. Que penser de la qualité de l'ajustement du modèle aux données au vu des figures 4, 5 et 6?

4. Déterminer le modèle marginal associé au modèle conditionnel posé en question 2.

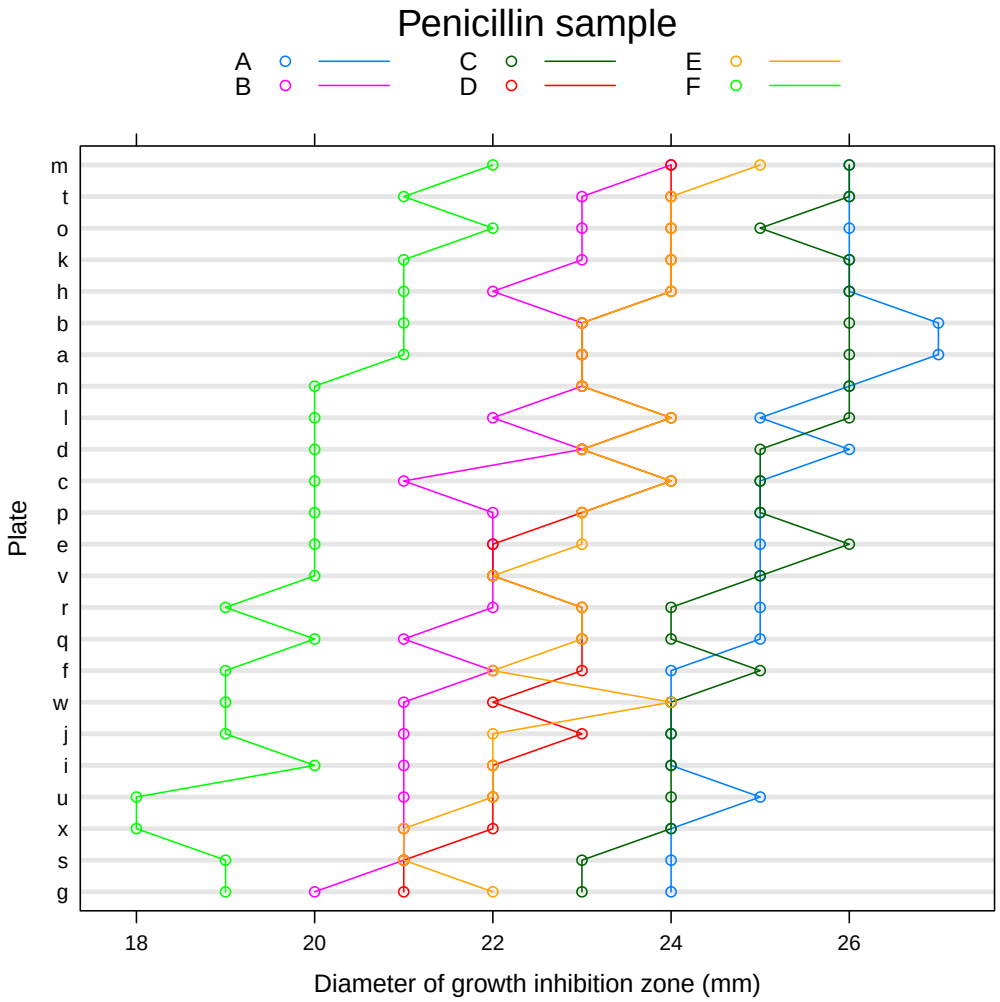

Figure 1: Tracé exploratoire relatif aux données `Penicillin`, question 1. de l'exercice .

Exercice 26.

1. Une étude est réalisée sur l'impact des radiations ionisantes sur les anomalies chromosomiques chez le rat. Pour cela, des rats sont exposés à des radiations ionisantes. Les rats sont regroupés par fratrie de taille respective 5, 5, 5, 6, 6, 6, 6, 6, 7, 7, 8 dans différentes cages. Les cages sont soumises quotidiennement pendant deux mois à des radiations ionisantes pendant une durée identique. Les différentes cages reçoivent chacune une dose différente de radiations ionisantes. Au terme de la période d'exposition, un

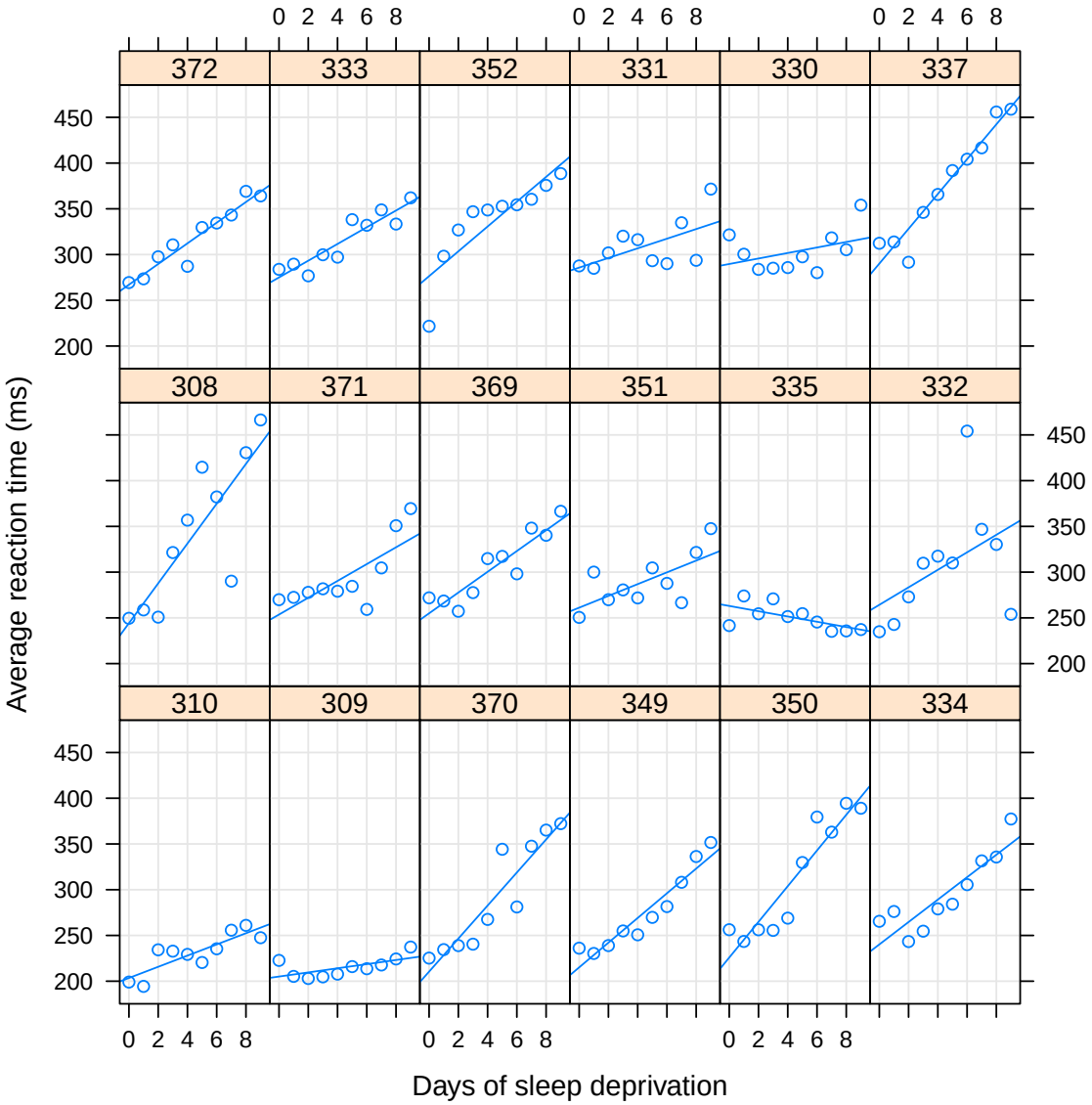

Figure 2: Tracé exploratoire relatif aux données `sleepstudy`, question 1. de l'exercice .

prélèvement sanguin est réalisé sur chaque rat. L'analyse de chaque prélèvement fournit le nombre d'anomalies chromosomiques relevées dans le prélèvement ainsi que le nombre de cellules analysées. Proposer un modèle adapté à cette étude. Comment tester l'impact des radiations ionisantes sur le nombre d'anomalies chromosomiques? Mettre en oeuvre. On note ED_{10} la dose de radiations ionisantes associée à une proportion d'anomalies chromosomiques dans un prélèvement égale à 10%. Comment estimer ED_{10}?

2. Une étude est réalisée chez le rat sur le lien entre l'absorption d'un neurotoxique et l'état d'excitation. Pour cela, on soumet des rats de même âge, sans lien de parenté, vivant dans différentes cages, à différentes doses de ce neurotoxique. Pour chaque rat, on mesure un score de sévérité d'excitation à $t = 0$ (juste avant l'administration du neurotoxique), à $t = 4h$ après l'administration du neurotoxique et à $t = 24h$ après l'administration du neurotoxique. Plus le score est élevé, plus le rat est excité. Les différentes doses administrées sont 0, 150, 500, 1500 et 5000 mg/kg. Chaque niveau de dose est aléatoirement attribué à 8 rats. Un tracé exploratoire est réalisé en figure 7. Proposer un modèle adapté à cette

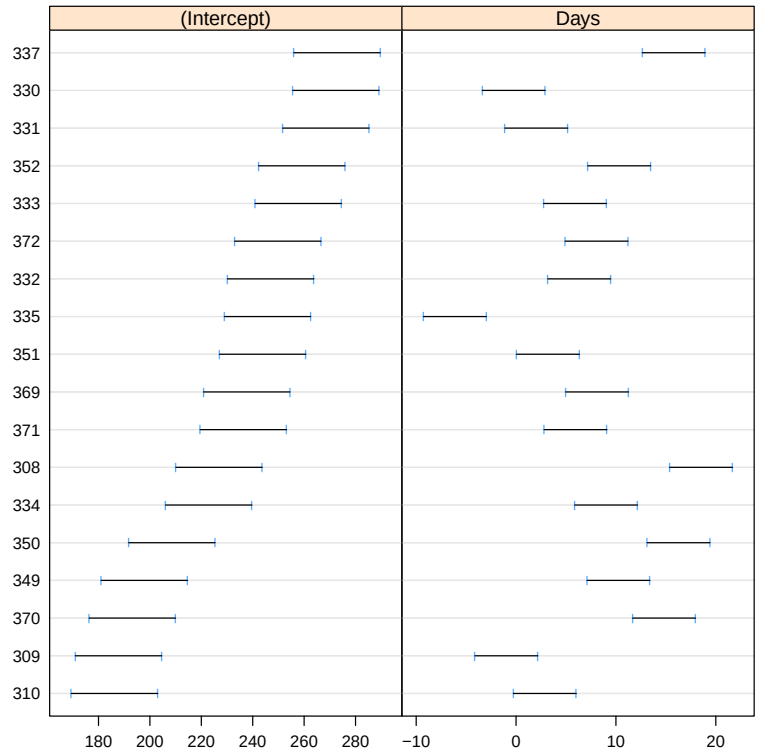

Figure 3: Tracé exploratoire relatif aux données `sleepstudy`, question 1. de l'exercice .

étude et interpréter ses paramètres.

Exercice 27.
Un épidémiologiste étudie la présence d'hypertension artérielle chez 50 familles. Au moins 3 membres d'une même famille sont inclus dans l'étude, constituant ainsi un échantillon de taille 204. L'épidémiologiste ajuste un modèle linéaire généralisé à effets mixtes sur ces données en utilisant le lien probit. La variable réponse code la présence ou l'absence d'hypertension artérielle. Sont inclus les prédicteurs suivants: sexe, âge, IMC (indice de masse corporelle), taux sanguin de potassium (exprimé en mg/L), consommation moyenne hebdomadaire d'alcool (exprimée en dL par kilo de poids corporel). Aucun terme d'interaction n'est inclus dans le modèle.

1. Ecrire l'équation et les hypothèses du modèle linéaire généralisé à effets mixtes utilisé par cet épidémiologiste.

2. Déterminer le modèle marginal.

3. Quel autre modèle aurait-il pu proposer?

Exercice 28.
L'étude suivante est réalisée chez 10 chiens. Un scanner est effectué pour chaque chien avec injection de produit de contraste à plusieurs reprises. A chaque fois est mesurée l'intensité

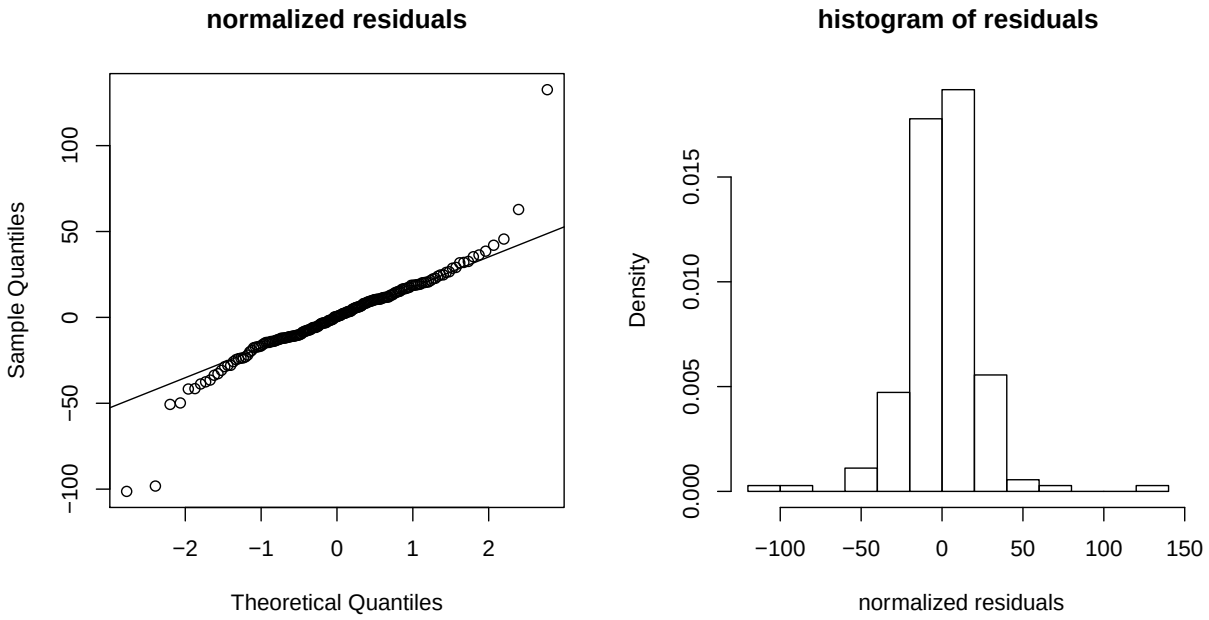

Figure 4: Tracé relatif aux résidus normalisés obtenus dans l'étude des données `sleepstudy`, question 3. de l'exercice .

moyenne des pixels correspondant au ganglion lymphatique de la région axillaire[1] droite ainsi que l'intensité moyenne des pixels correspondant au ganglion lymphatique de la région axillaire gauche. Les données recueillies sont disponibles dans le jeu de données `Pixel` du package `nlme`. Une brève exploration graphique de ces données est effectuée avec le code ci-dessous. Les figures correpondantes sont présentées en figures 8 et 9.

```
> library(nlme)
> data(Pixel)
> attach(Pixel)
> help(Pixel)
> str(Pixel)
Classes ‘nmGroupedData’, ‘groupedData’ and ’data.frame’: 102 obs. of  4 variables:
 $ Dog  : Factor w/ 10 levels "1","10","2","3",..: 1 1 1 1 1 1 1 3 3 3 ...
 $ Side : Factor w/ 2 levels "L","R": 2 2 2 2 2 2 2 2 2 2 ...
 $ day  : num  0 1 2 4 6 10 14 0 1 2 ...
 $ pixel: num  1046 1044 1043 1050 1045 ...
 - attr(*, "formula")=Class ’formula’  language pixel ~ day | Dog/Side
  .. ..- attr(*, ".Environment")=<environment: R_GlobalEnv>
 - attr(*, "formulaList")=List of 2
  ..$ Dog :Class ’formula’  language ~Dog
  .. .. ..- attr(*, ".Environment")=<environment: R_GlobalEnv>
  ..$ Side:Class ’formula’  language ~Side
  .. .. ..- attr(*, ".Environment")=<environment: R_GlobalEnv>
 - attr(*, "labels")=List of 2
  ..$ x: chr "Time post injection"
  ..$ y: chr "Pixel intensity"
```

[1]La région axillaire chez le chien correspond à une zone du thorax proche des pattes avant.

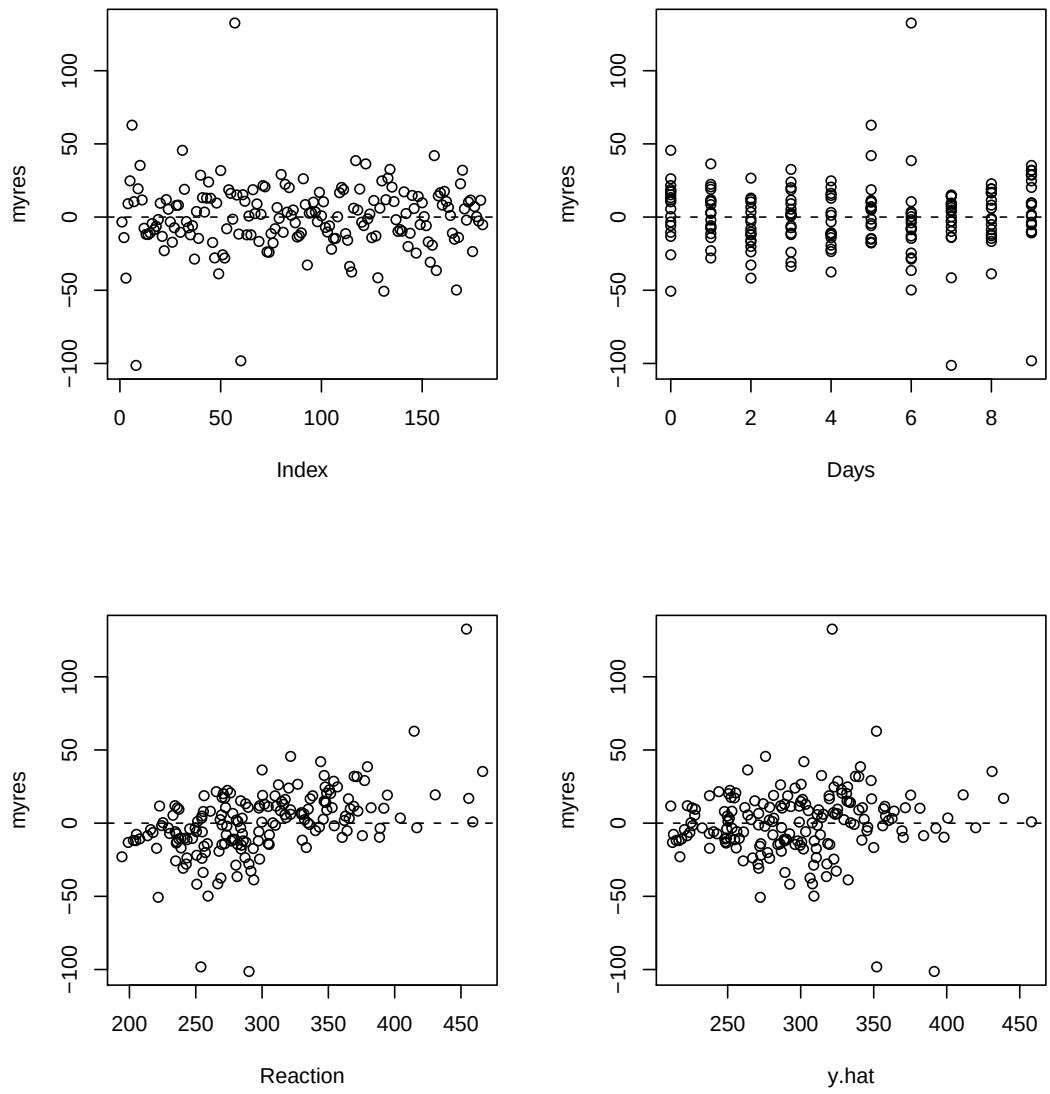

Figure 5: Tracé relatif aux résidus normalisés obtenus dans l'étude des données `sleepstudy`, question 3. de l'exercice .

```
 - attr(*, "units")=List of 1
  ..$ x: chr "(days)"
 - attr(*, "order.groups")=List of 2
  ..$ Dog : logi TRUE
  ..$ Side: logi TRUE
 - attr(*, "FUN")=function (x)
  ..- attr(*, "source")= chr "function (x) max(x, na.rm = TRUE)"
>
> library(lattice)
> library(lme4)
>
> xyplot(pixel ~ day | Dog,
+        data=Pixel,
```

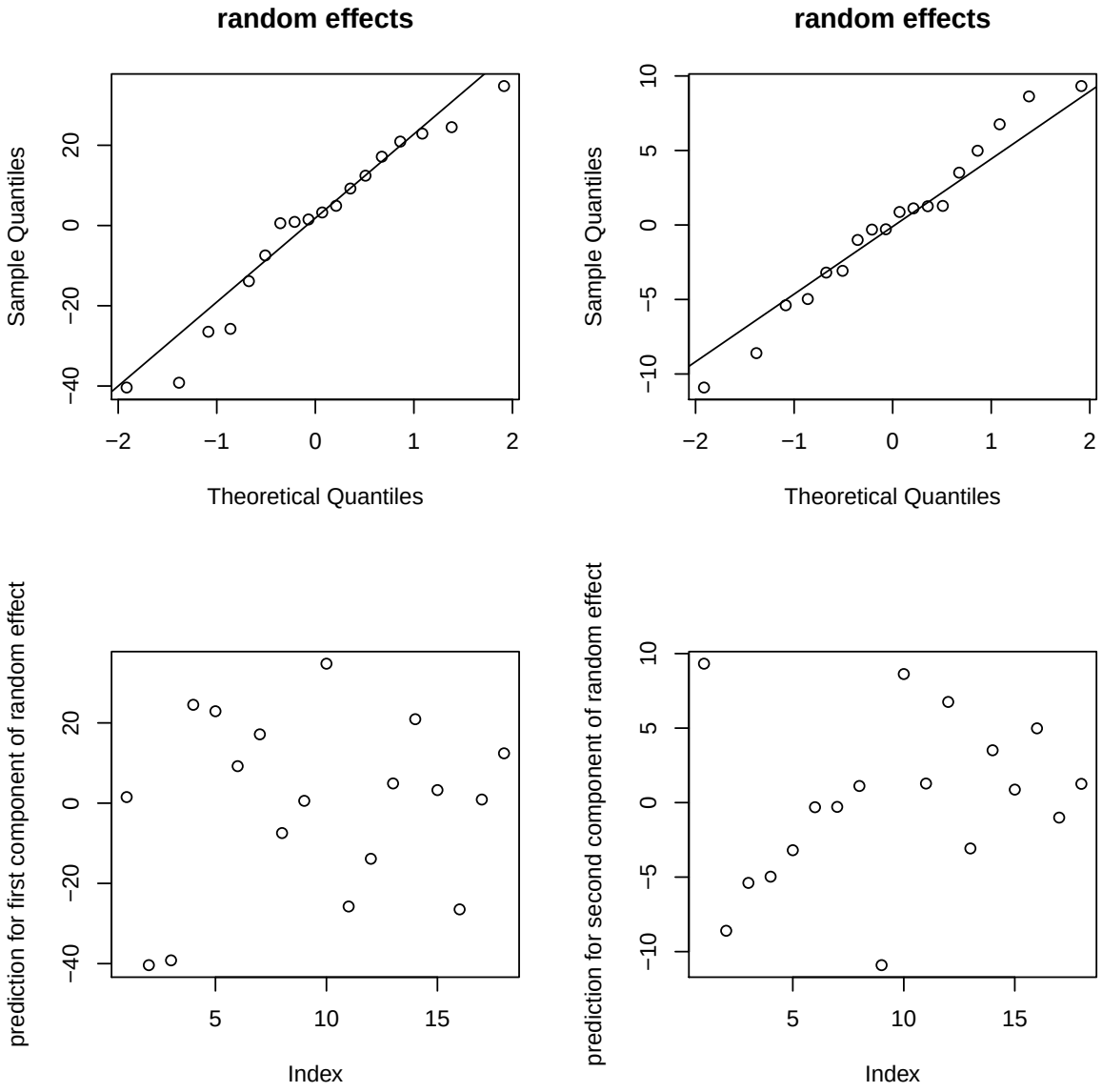

Figure 6: Tracé relatif aux effets aléatoires prédits obtenus dans l'étude des données `sleepstudy`, question 3. de l'exercice .

```
+        groups=Side,
+        type = c("g","p","a"),
+        xlab = "day",
+        ylab = "mean pixel intensity",
+        aspect = "xy")
>
> manyfitsL<-lme4::lmList(pixel ~ day+ I(day^2)|Dog,subset=Side=="L",data=Pixel)
> x11()
> plot(confint(manyfitsL,pooled=TRUE),order=1,main="left side")
> manyfitsR<-lme4::lmList(pixel ~ day+ I(day^2)|Dog,subset=Side=="R",data=Pixel)
> x11()
> plot(confint(manyfitsR,pooled=TRUE),order=1,main="right side")
```

1. Dans un premier temps, le modèle présenté dans le code ci-dessous est ajusté. Les tracés

correspondant sont présentés en figures 10, 11 et 12.

```
> myfit<-lme4::lmer(pixel ~ day+I(day^2)+(day+I(day^2)|Dog/Side),data=Pixel)
Warning messages:
1: In optwrap(optimizer, devfun, getStart(start, rho$lower, rho$pp),  :
  convergence code 1 from bobyqa:
bobyqa -- maximum number of function evaluations exceeded
2: In checkConv(attr(opt, "derivs"), opt$par, ctrl = control$checkConv,  :
  Model failed to converge with max|grad| = 1.11639 (tol = 0.002, component 1)
> summary(myfit)
Linear mixed model fit by REML ['lmerMod']
Formula: pixel ~ day + I(day^2) + (day + I(day^2) | Dog/Side)
   Data: Pixel

REML criterion at convergence: 806.1

Scaled residuals:
     Min       1Q   Median       3Q      Max
-2.61954 -0.44148 -0.00403  0.45481  2.94741

Random effects:
 Groups   Name        Variance  Std.Dev. Corr
 Side:Dog (Intercept)  32.14556  5.6697
          day           7.23020  2.6889   0.97
          I(day^2)      0.01240  0.1114  -1.00 -0.99
 Dog      (Intercept) 484.70472 22.0160
          day           6.30823  2.5116   0.02
          I(day^2)      0.01293  0.1137  -0.63 -0.78
 Residual              62.61584  7.9130
Number of obs: 102, groups:  Side:Dog, 20; Dog, 10

Fixed effects:
             Estimate Std. Error t value
(Intercept) 1071.41819    7.56368  141.65
day            6.27295    1.15344    5.44
I(day^2)      -0.36195    0.05195   -6.97

Correlation of Fixed Effects:
         (Intr) day
day      -0.061
I(day^2) -0.358 -0.867
convergence code: 1
Model failed to converge with max|grad| = 1.11639 (tol = 0.002, component 1)

>
> myres<-residuals(myfit,scaled=T)
> x11()
> qqnorm(myres,main="residuals")
> qqline(myres)
> shapiro.test(myres)
```

```
Shapiro-Wilk normality test
data:  myres
W = 0.95422, p-value = 0.001397

> x11()
> par(mfrow=c(1,2))
> plot(day,myres)
> abline(h=0,lty=2)
> plot(pixel,myres)
> abline(h=0,lty=2)
>
> xyplot(predict(myfit,type="response") ~ day | Dog,
+        data=Pixel,
+        groups=Side,
+        type = c("g","p","a"),
+        xlab = "day",
+        ylab = "predicted mean pixel intensity",
+        aspect = "xy")
```

Ecrire les équations et hypothèses définissant ce premier modèle ajusté après avoir introduit les notations nécessaires à la formalisation du problème.

2. Les auteurs Pinheiro et Bates proposent le modèle ajusté dans le code qui suit. Les tracés correspondant sont présentés en figures 13, 14 et 15.

```
> myfit_red<-lme4::lmer(pixel~day+I(day^2)+(day|Dog)+(1|Dog:Side),data=Pixel)
> summary(myfit_red)
Linear mixed model fit by REML ['lmerMod']
Formula: pixel ~ day + I(day^2) + (day | Dog) + (1 | Dog:Side)
Data: Pixel

REML criterion at convergence: 825.2

Scaled residuals:
     Min       1Q   Median       3Q      Max
-2.82906 -0.44918  0.02555  0.55722  2.75196

Random effects:
 Groups   Name        Variance Std.Dev. Corr
 Dog:Side (Intercept) 283.055  16.824
 Dog      (Intercept) 804.853  28.370
          day           3.399   1.844   -0.55
 Residual              80.813   8.990
Number of obs: 102, groups:  Dog:Side, 20; Dog, 10

Fixed effects:
             Estimate Std. Error t value
(Intercept) 1073.33914   10.17169  105.52
day            6.12960    0.87932    6.97
I(day^2)      -0.36735    0.03395  -10.82
```

```
Correlation of Fixed Effects:
          (Intr) day
day       -0.517
I(day^2)  0.186 -0.668
>
> myres_red<-residuals(myfit_red,scaled=T)
> x11()
> qqnorm(myres_red,main="residuals")
> qqline(myres_red)
> shapiro.test(myres_red)
Shapiro-Wilk normality test
data:  myres_red
W = 0.97197, p-value = 0.02873

> x11()
> par(mfrow=c(1,2))
> plot(day,myres_red)
> abline(h=0,lty=2)
> plot(pixel,myres_red)
> abline(h=0,lty=2)
```

Ecrire les équations et hypothèses définissant ce deuxième modèle ajusté.

3. Sur lequel de ces deux modèles se porterait votre choix et pourquoi? Expliquer les résultats obtenus pour le chien n^o 10.

4. Quel autre type de modèle aurait-on pu écrire? Ecrire l'alternative correspondant au premier modèle présenté, puis au second modèle présenté. Qu'est-ce que ce nouveau modèle ne permet pas ou difficilement de modéliser ici?

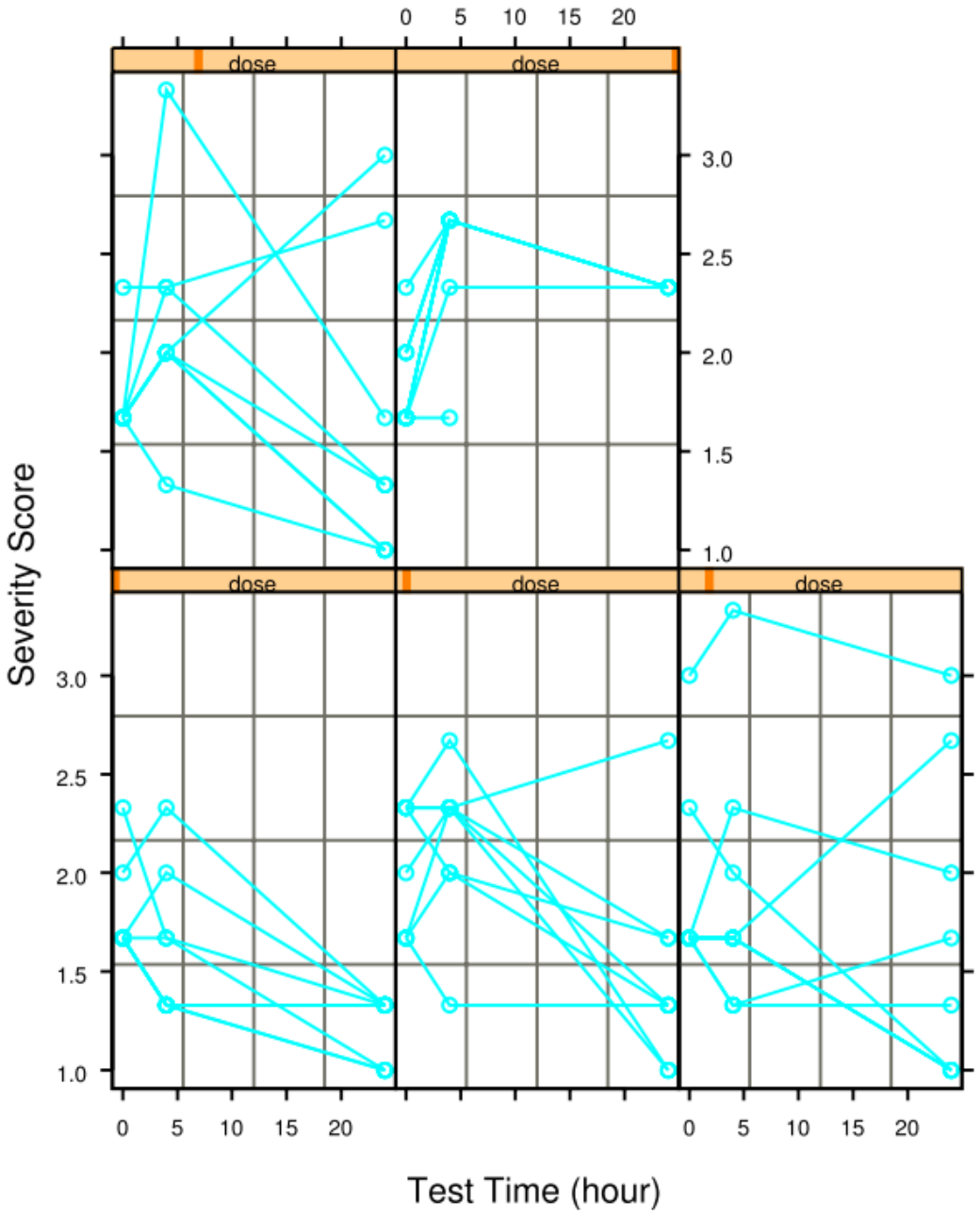

Figure 7: Tracé exploratoire relatif aux données de la question 2. de l'exercice . Le score de sévérité est tracé en fonction du temps pour chaque rat par groupe de dose. Les doses augmentent de case en case de gauche à droite et de bas en haut. Dans le groupe de rats recevant la dose la plus élevée, quatre rats sont morts avant 24h.

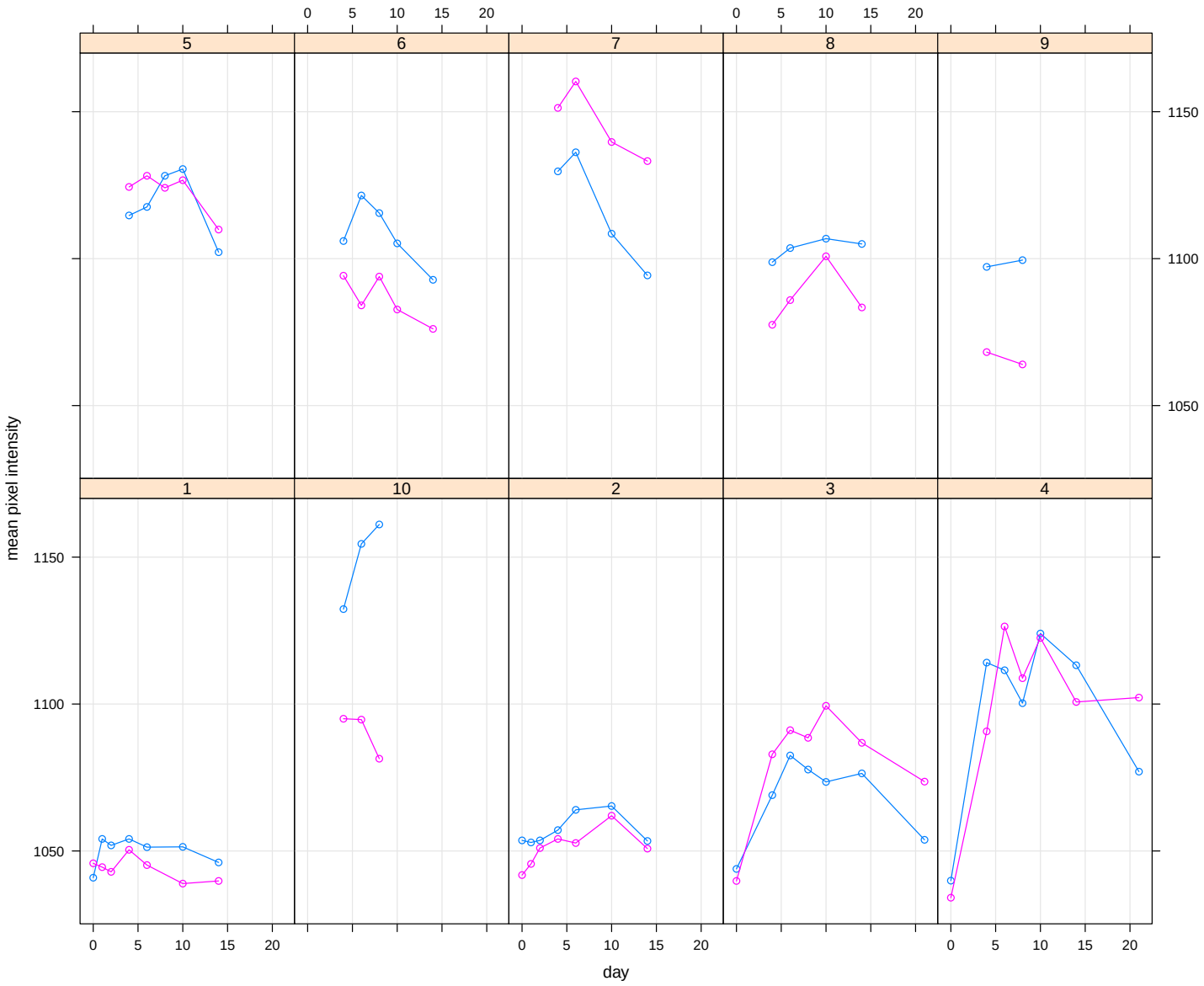

Figure 8: Tracé exploratoire pour l'exercice

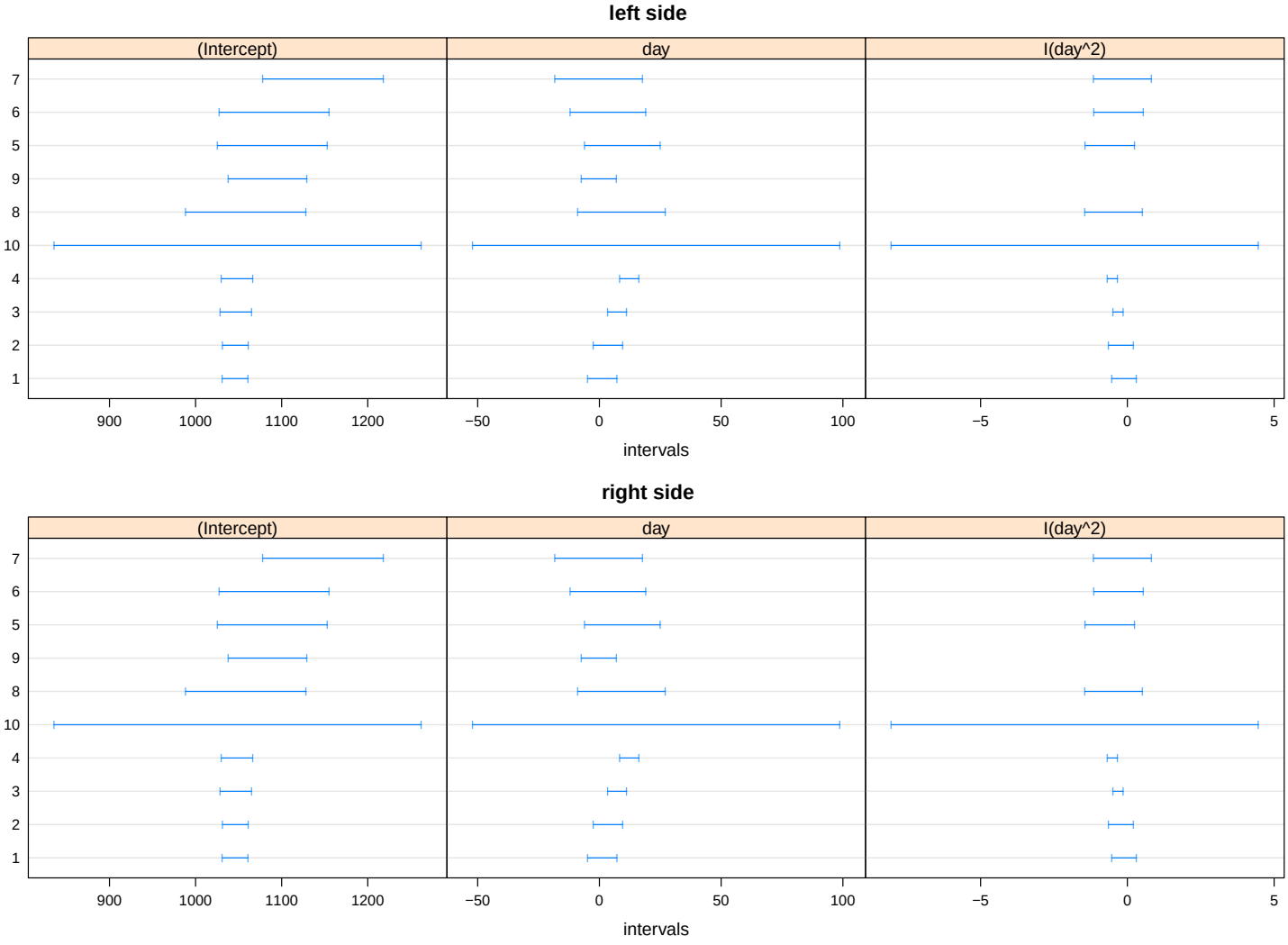

Figure 9: Tracé exploratoire (fin) pour l'exercice

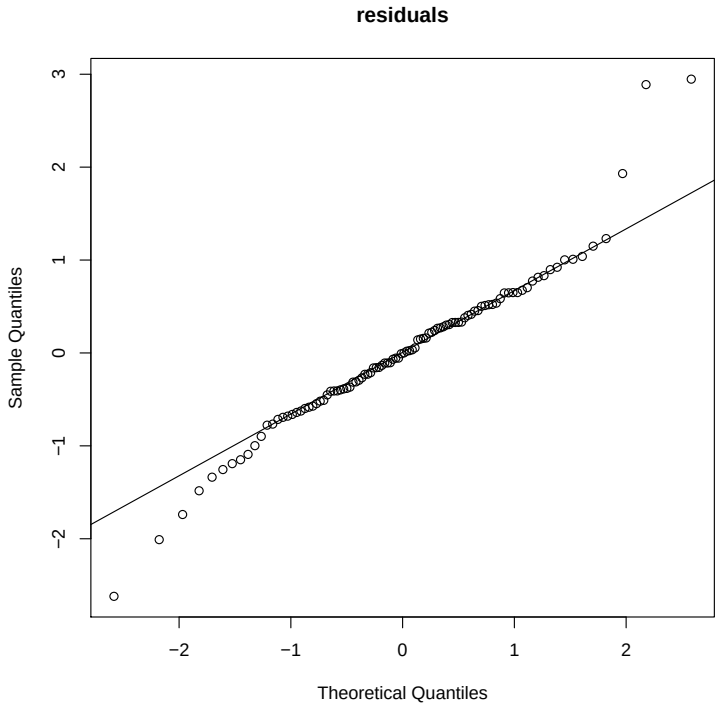

Figure 10: Qqplot obtenu pour le premier modèle présenté dans l'exercice

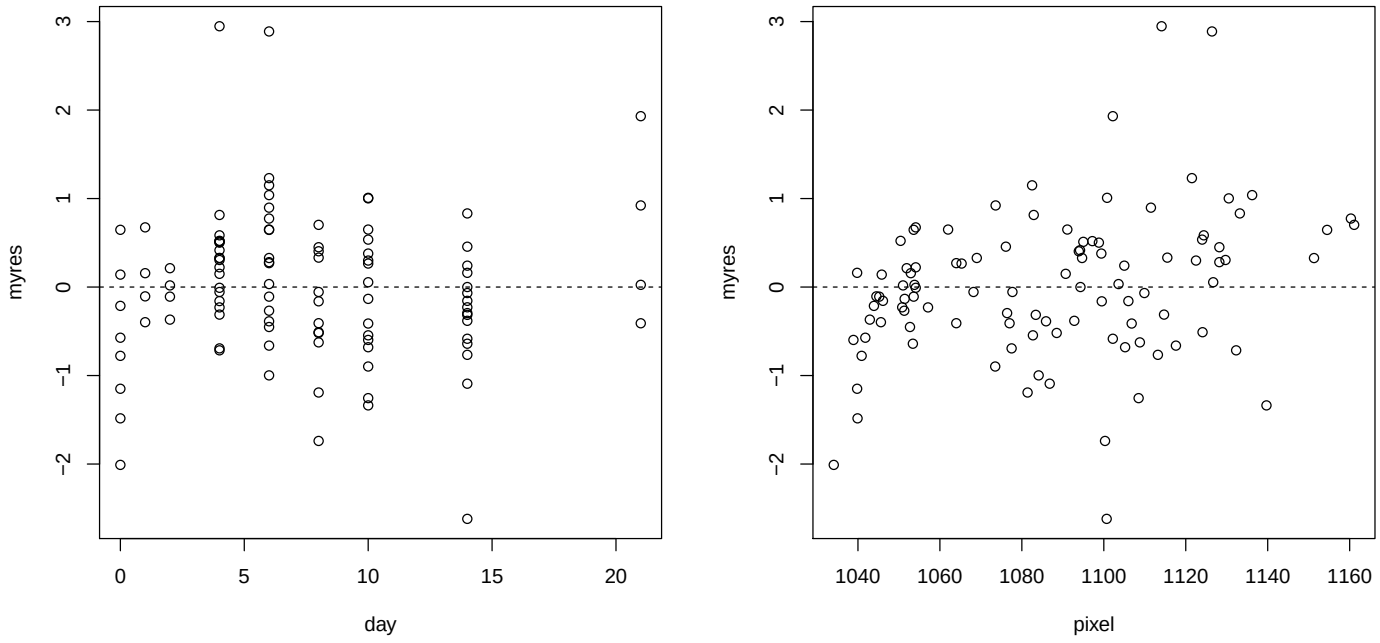

Figure 11: Tracé obtenu pour le premier modèle présenté dans l'exercice

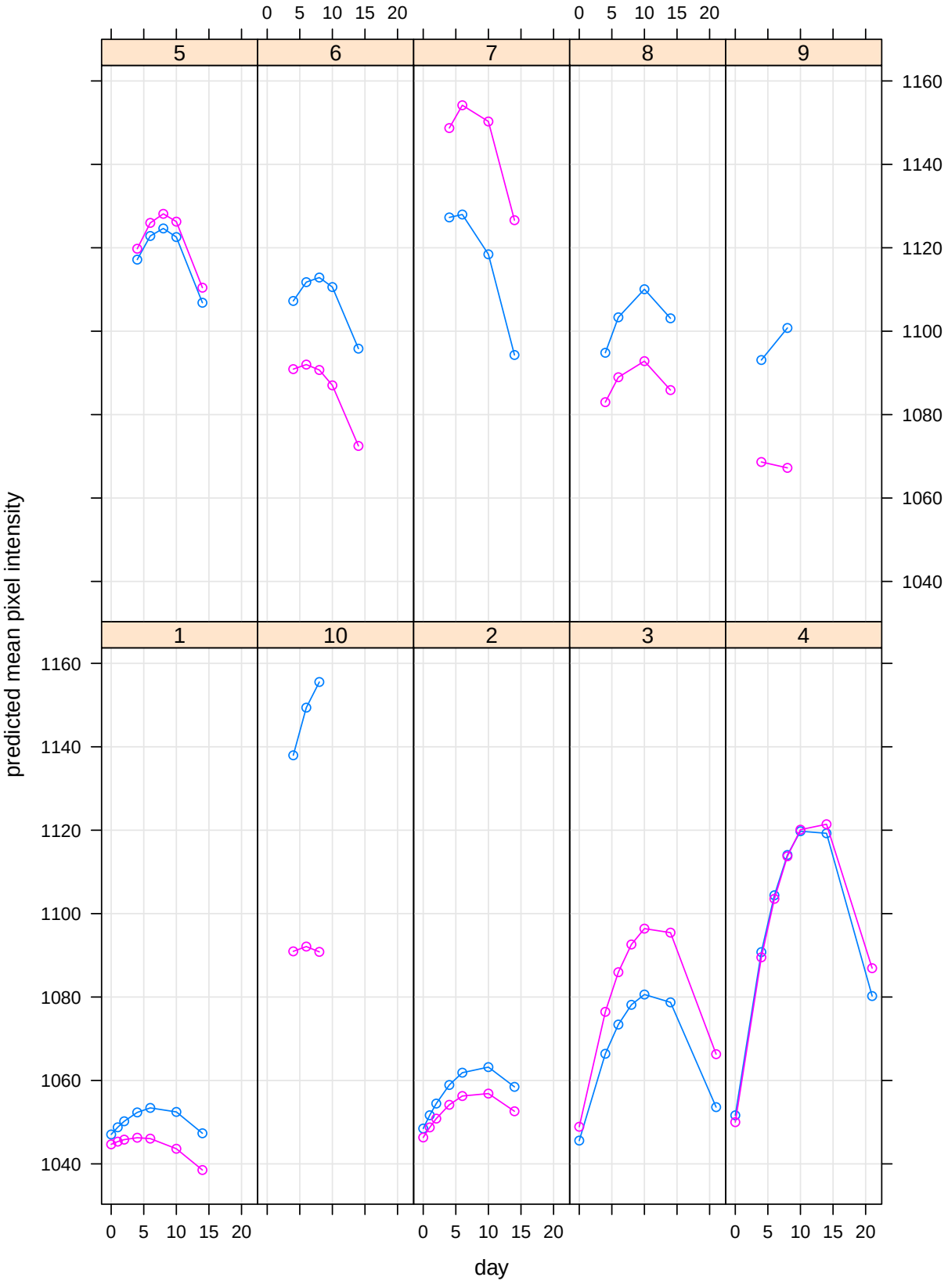

Figure 12: Tracé obtenu pour le premier modèle présenté dans l'exercice

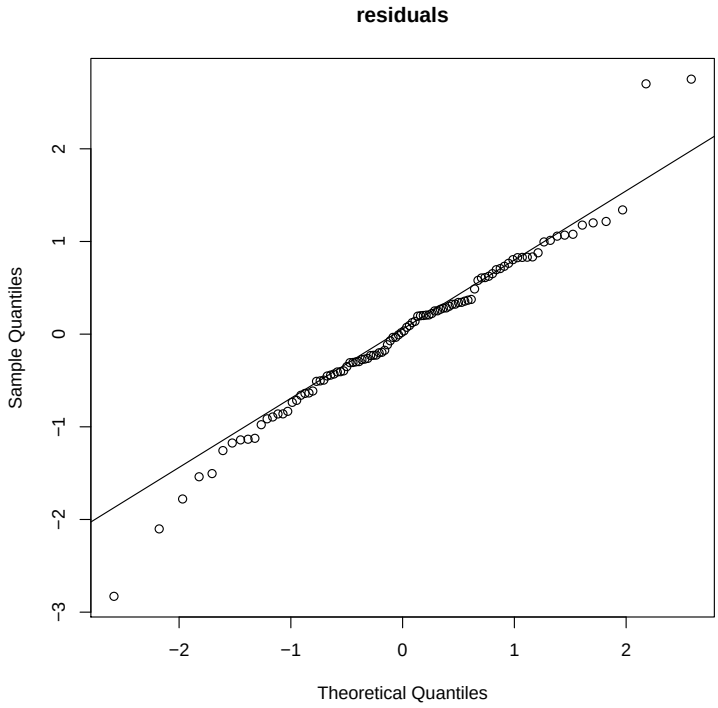

Figure 13: Qqplot obtenu pour le deuxième modèle présenté dans l'exercice

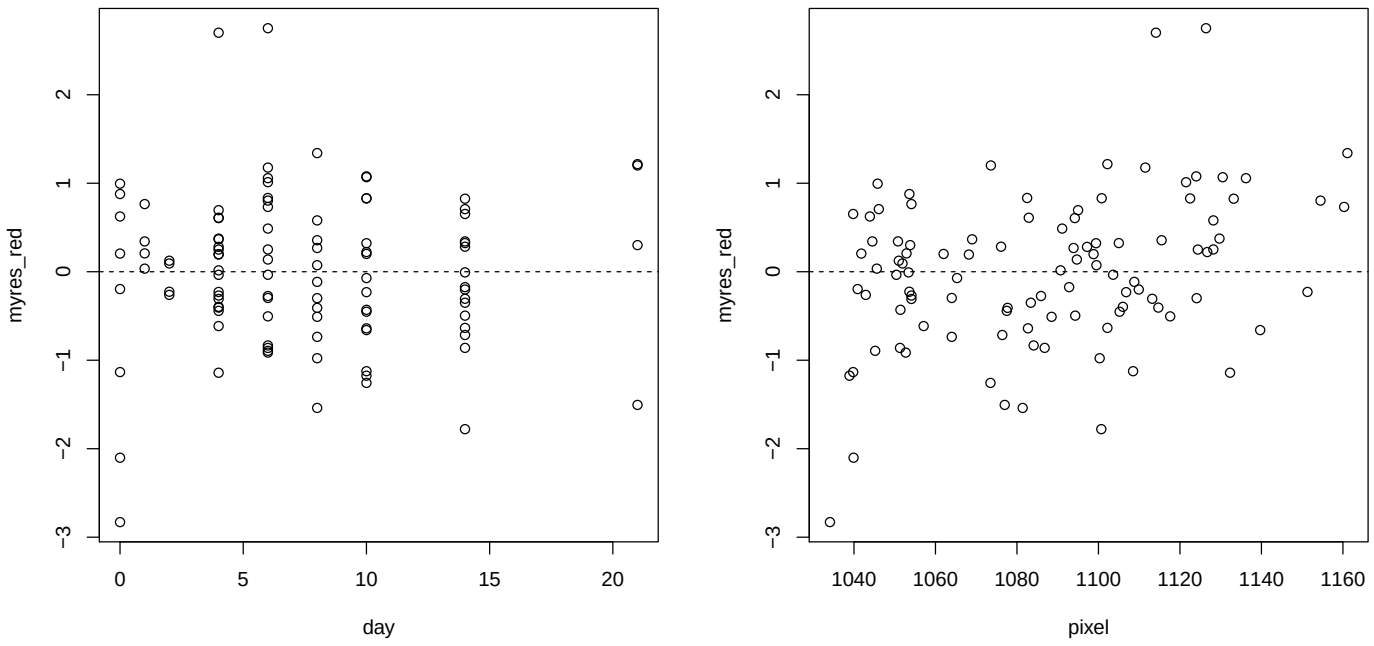

Figure 14: Tracé obtenu pour le deuxième modèle présenté dans l'exercice

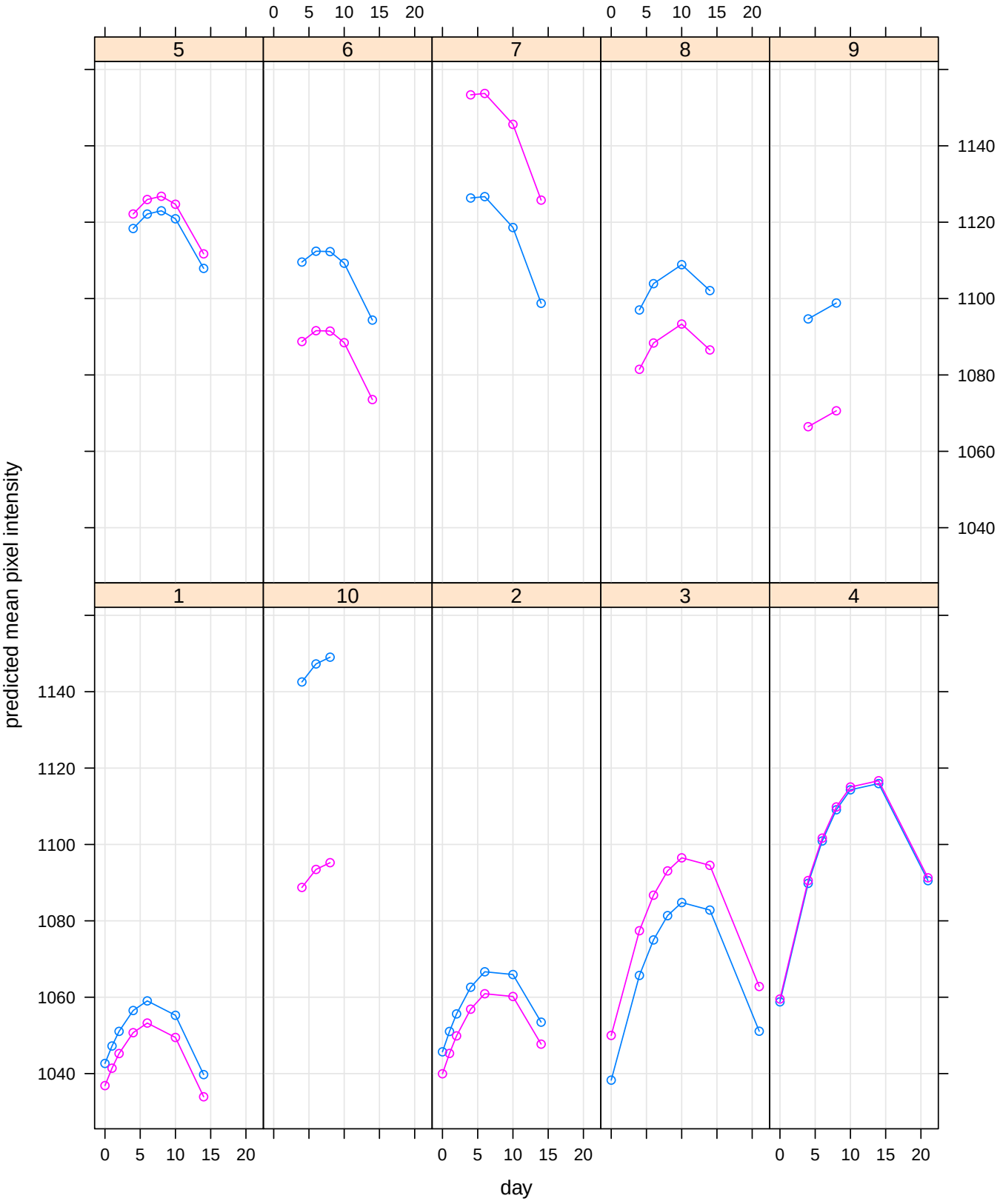

Figure 15: Tracé obtenu pour le deuxième modèle présenté dans l'exercice