Data Analysis Report

Aritz Lizoain

April 2022

Abstract

A Generalized Linear Mixed Model (GLMM) is fitted in order to study the correlated observations in longitudinal data of patients suffering from schizophrenia. The model parameter estimates reveal that 'male' patients evolve less favorably than 'female' patients, while age does not have a significant effect on the evolution of prevalence of thought disorders.

1 Introduction

The dataset contains longitudinal data of 86 patients suffering from schizophrenia, repeatedly examined over a year. Each observation consists of the following information: patient ID number (ID), indicator of thought disorders (Y), age (AGE), gender ($GENDER$), and month of observation ($MONTH$). The purpose of the analysis is to determine the relevance of different patient characteristics in the evolution of thought disorders.

2 Statistical Modelling

2.1 Exploratory Data Analysis

Before starting to explore the dataset, a quick filtering is done; patients that do not contain observations for all 12 months are removed. This filtering results in a dataset reduced to 69 individuals.

The general evolution of thought disorders over time can be uncovered by analyzing Y as a function of $MONTH$ (see Figure 1). A clear difference is found between the distribution of observations with presence and observations with absence of thought disorders; it is more common to have thought disorders during the first months than the last. The correlation value of -0.45 indicates the same behavior; the later in time, the lower the thought disorder proportion.

The wide variety of thought disorder evolution patterns among individuals poses a great challenge for outlier detection. However, after discerning the distribution discrepancies of observations with and without thought disorders, it seems reasonable to expect any patient to have less thought disorders in the last 6 months than in the first 6 months. When the number of months with thought disorders in the first 6 months are plotted against the number of months with thought disorders in the last 6 months, two potential outliers are found (see Figure 2). Removing these two patients from the dataset could be a valid option. Nevertheless, due to the aforementioned wide variety of patterns found in the dataset, it is conservatively decided to not remove them.

In order to get the first insights regarding the influence of age and gender in the evolution of thought disorders, means of the thought disorder indicator Y for each age group (see Figure 3), and for each gender group (see Figure 4), are plotted against the time variable $MONTH$. The proportion of patients having no thought disorders is shown in the background.

Figure 3 shows little difference in the evolution of thought disorders among 'young' and 'old' patients, with the overall level slightly higher for 'old' patients. A more striking difference is observed between gender groups in Figure 4, where the overall level is higher for 'male' patients, and the decrease over time clearly greater for 'female' patients.

Figure 1. Boxplot comparing the distribution of observations with presence of thought disorders, and observations with absence of thought disorders. The correlation value corresponds to the correlation between the indicator of thought disorders Y, and the month of observation $MONTH$.

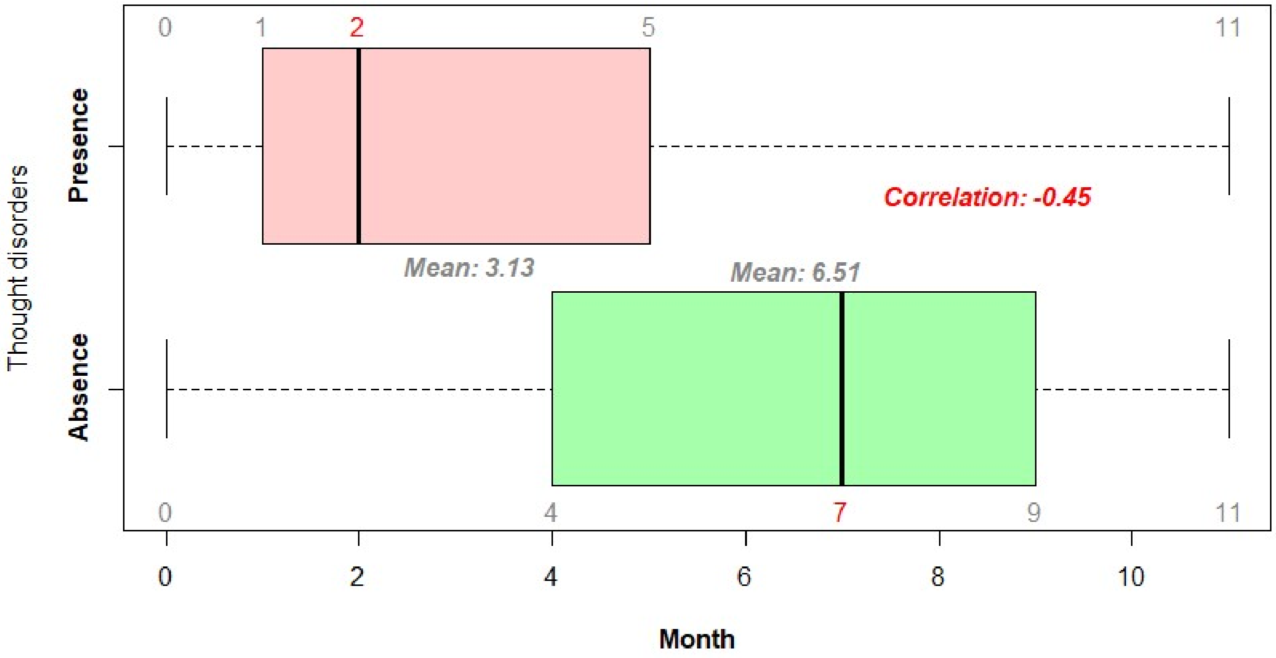


Figure 2. Scatterplot of number of months with thought disorders in the first 6 months against the last 6 months. The red line represents the limit of expected behavior, i.e. having less months with thought disorders in the last 6 months. The points in red are potential outliers. For visualization purposes, the points are not exactly located in the integer coordinates, but the values do correspond to integers.

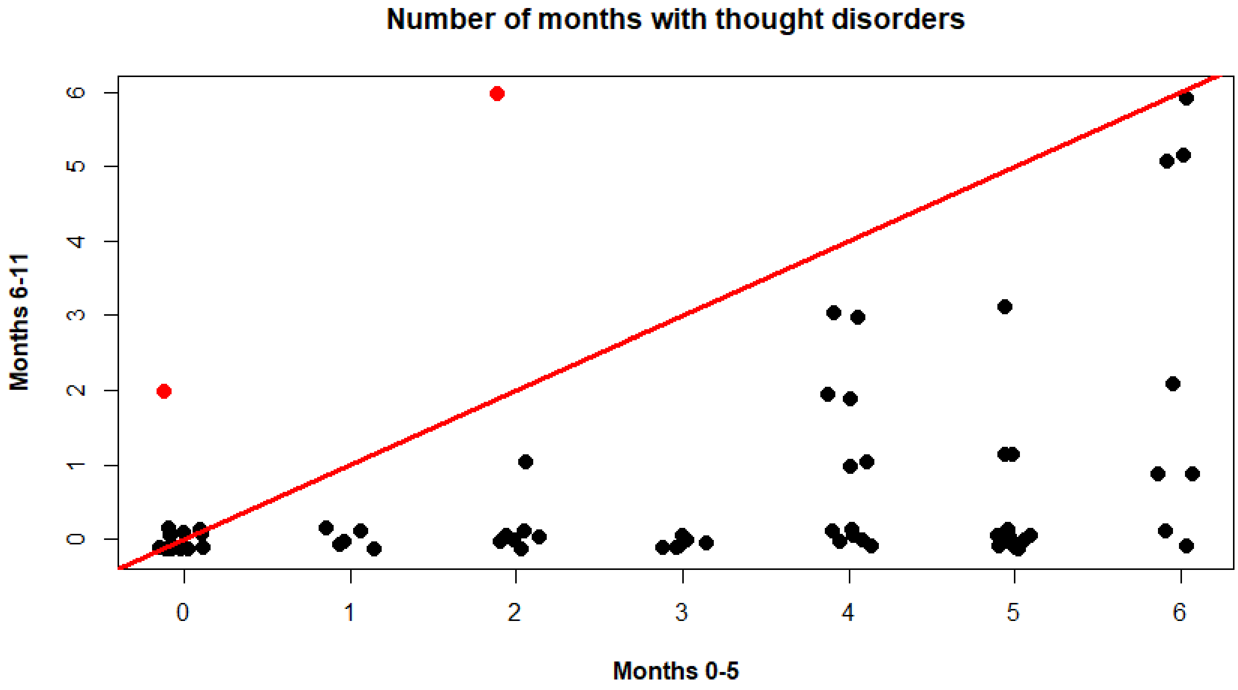


Figure 3. Monthly mean values of the thought disorder indicator Y for each age group. The background shows the proportion of patients having no thought disorders.

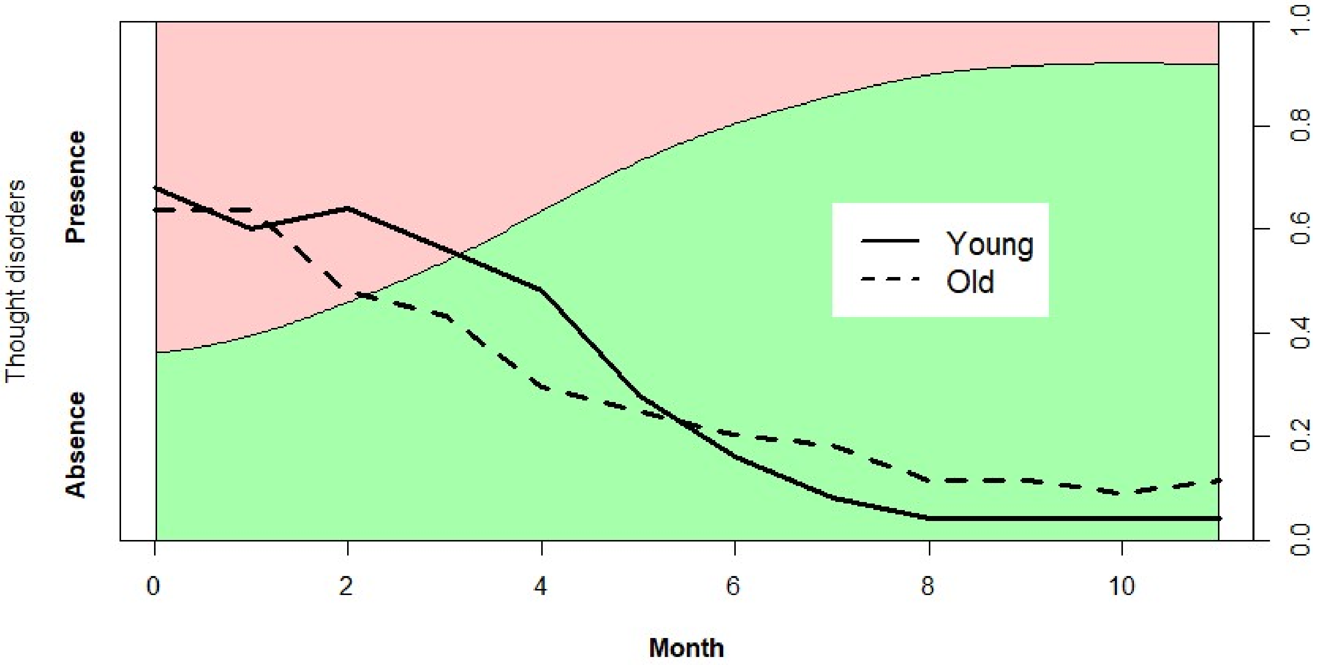
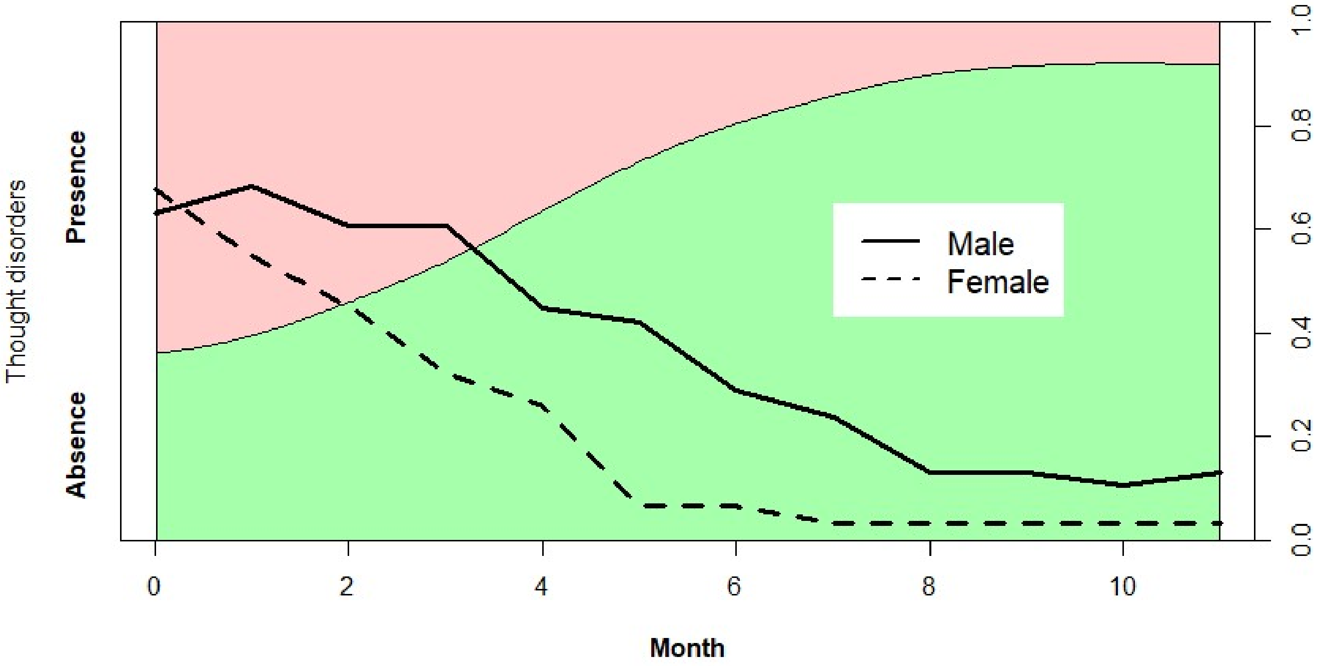


Figure 4. Monthly mean values of the thought disorder indicator Y for each gender group. The background shows the proportion of patients having no thought disorders.



2.2 Model formulation

The correlated nature of longitudinal data prevents the employment of Generalized Linear Models (GLM), which assume independence of observations. In this case study, where individual patients are of interest, there exists a reason to include individual-level randomness. Therefore, a Generalized Linear Mixed Model (GLMM) is used, which comprises fixed and random effects.

Let Y_{ij} denote the response variable associated with a subject-specific intercept γ_i as random element, and a covariate pattern $x_{ij} = (x_{ij1}, x_{ij2}, x_{ij3}, x_{ij1} \cdot x_{ij2}, x_{ij1} \cdot x_{ij3})^T$, with $x_{ij1} = MONTH_{ij}$, $x_{ij2} = \mathbb{1}_{\{GENDER_{ij}=\text{"}male\text{"}\}}$, and $x_{ij3} = \mathbb{1}_{\{AGE_{ij}=\text{"}young\text{"}\}}$, where $\mathbb{1}_{\{.\}}$ denotes the indicator function. Since the response variable is binary, a binomial model with logit link function is chosen.

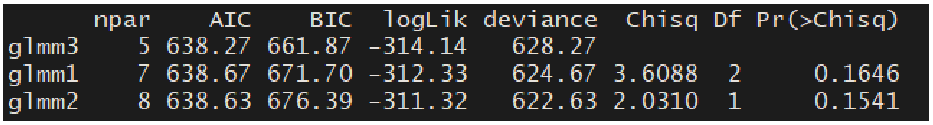
The model ($glmm1$) can be written as

$$Y_{ij} \sim Bernoulli(p_{ij})$$
$$log(\frac{p_{ij}}{1-p_{ij}}) = \beta_0 + \beta_1 \cdot x_{ij1} + \beta_2 \cdot x_{ij2} + \beta_3 \cdot x_{ij3} + \beta_4 \cdot x_{ij1} \cdot x_{ij2} + \beta_5 \cdot x_{ij1} \cdot x_{ij3} + \gamma_i \tag{1}$$

2.3 Model selection

Besides the model described by eq.(1), two more models are fitted. On the first one ($glmm2$), an interaction between $GENDER$ and AGE, i.e. $x_{ij2} \cdot x_{ij3}$, is added. On the second one ($glmm3$) instead, the model is simplified by removing the AGE and $GENDER$ factors. All three models are subsequently compared performing a difference in deviance test (see Figure 5) to assess the relevance of the additional variables.

Figure 5. Difference in deviance test of the model described by eq.(1) ($glmm1$), a model with an additional interaction parameter $GENDER \cdot AGE$ ($glmm2$), and a simplified model without AGE and $GENDER$ ($glmm3$).

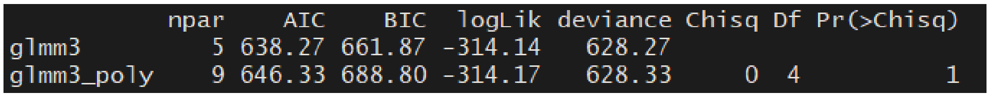


```
      npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
glmm3    5 638.27 661.87 -314.14   628.27                     
glmm1    7 638.67 671.70 -312.33   624.67 3.6088  2     0.1646
glmm2    8 638.63 676.39 -311.32   622.63 2.0310  1     0.1541
```

The test reveals that the additional variables do not lead to a significantly improved fit, thereby proclaiming the simplest model ($glmm3$) to be the best compromise between model complexity and ability to fit the data.

Moreover, the significance of various polynomial components is tested. In order to assess the relevance of the quadratic and cubic effect, a new model ($glmm3\_poly$) is created with polynomials up to degree 3 for the interaction terms $MONTH \cdot AGE$ and $MONTH \cdot GENDER$. The new model ($glmm3\_poly$) is then compared to the model previously determined as the best ($glmm3$), performing the difference in deviance test (see Figure 6).

Figure 6. Difference in deviance test of the best model ($glmm3$), and a model with additional quadratic and cubic interaction terms ($glmm3\_poly$).



```
           npar    AIC    BIC  logLik deviance Chisq Df Pr(>Chisq)
glmm3         5 638.27 661.87 -314.14   628.27                    
glmm3_poly    9 646.33 688.80 -314.17   628.33     0  4          1
```

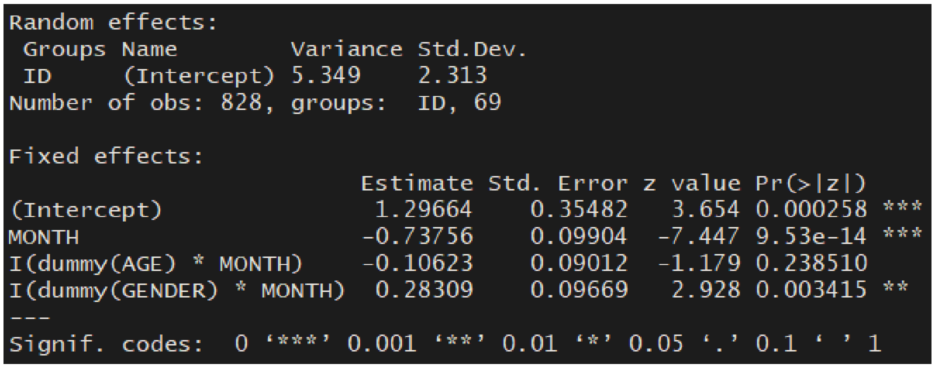
This second test reveals that the additional quadratic and cubic interaction terms do not lead to a significantly improved fit.

2.4 Parameter estimation

The model is fitted using the function 'glmer' from the R package $lme4$, obtaining the fixed effect parameter estimates shown by Figure 7. The selected model ($glmm3$) can be written as

$$log(\frac{p_{ij}}{1-p_{ij}}) = 1.297 - 0.738 \cdot x_{ij1} + 0.283 \cdot x_{ij1} \cdot x_{ij2} - 0.106 \cdot x_{ij1} \cdot x_{ij3} + \gamma_i \tag{2}$$

Figure 7. Random effects information and fixed effect coefficients of the fitted model (*glmm*3).



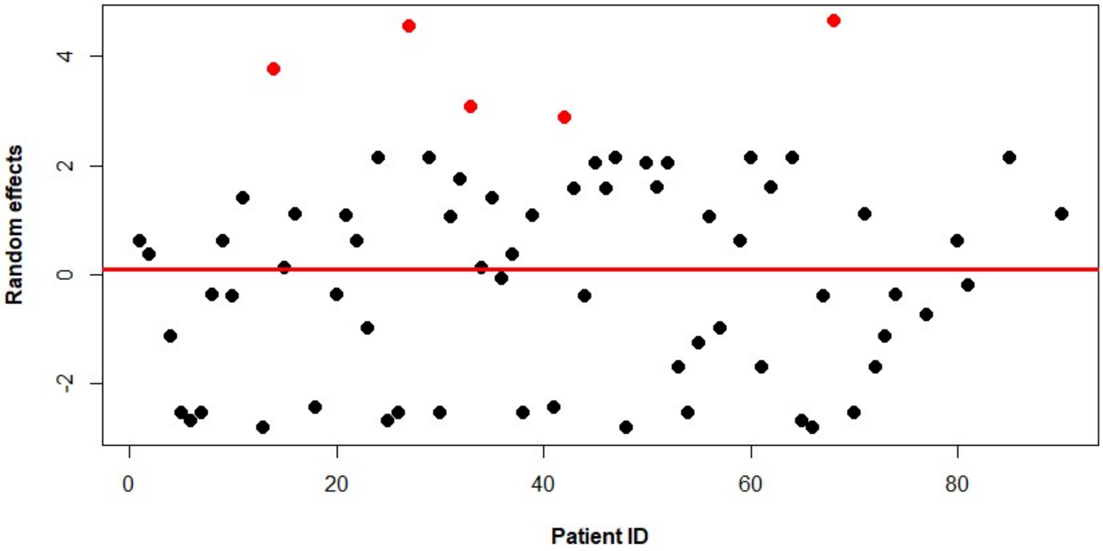
```
Random effects:
 Groups Name        Variance Std.Dev.
 ID     (Intercept) 5.349    2.313   
Number of obs: 828, groups:  ID, 69

Fixed effects:
                          Estimate Std. Error z value Pr(>|z|)    
(Intercept)                1.29664    0.35482   3.654 0.000258 ***
MONTH                     -0.73756    0.09904  -7.447 9.53e-14 ***
I(dummy(AGE) * MONTH)     -0.10623    0.09012  -1.179 0.238510    
I(dummy(GENDER) * MONTH)   0.28309    0.09669   2.928 0.003415 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

2.5 Model Validation

The first step to evaluate model adequacy is to verify the model assumptions. GLMMs assume that the random effects have mean 0, have constant variance, are uncorrelated, and come from a normal distribution. The random effects are plotted in Figure 8, with a mean value of 0.109, and several particular points colored in red. These points correspond to very unique individuals. The largest positive value ($ID = 68$), for instance, corresponds to a patient who had thought disorders every single month. The largest negative values, on the other hand, correspond to patients who did not have a single thought disorder in the whole year.

Figure 8. Random effects from the fitted model (*glmm*3). The five largest values are colored in red, and the red line corresponds to the mean value of 0.109.

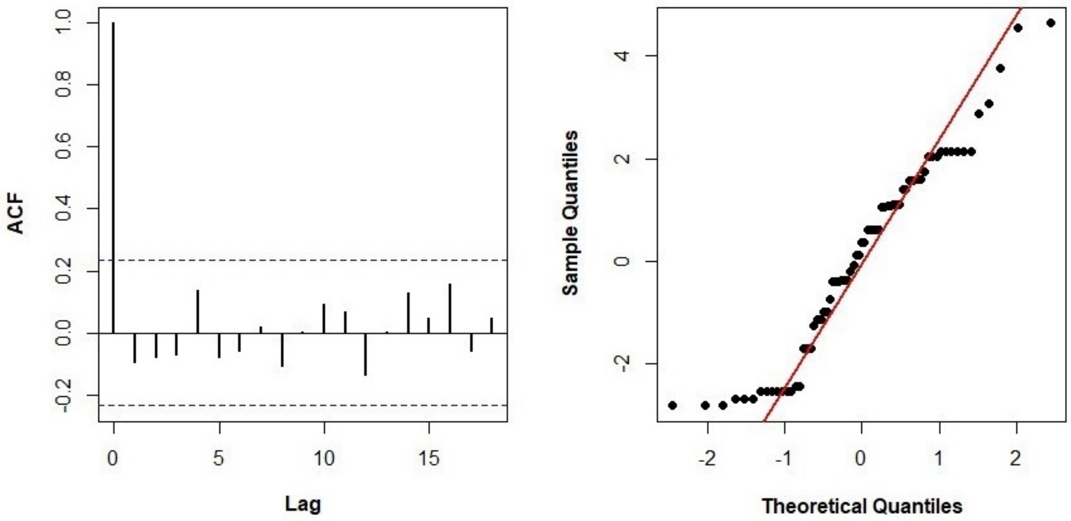


The autocorrelation function plot (ACF) of the random effects shows no correlation (see Figure 9 Left). Regarding the normality assumption, the left tail of the QQ-plot is expected due to the relatively large number of patients with no thought disorders throughout the year, and the distribution is considered to be approximately normal (see Figure 9 Right).

In order to evaluate the goodness of fit, the residuals are analyzed. Nonetheless, ”it has been shown that ordinary residuals in the analysis of longitudinal data are correlated and are not normally distributed”[1], and consequently unsuitable for analysis. For that reason, the residual analysis is done

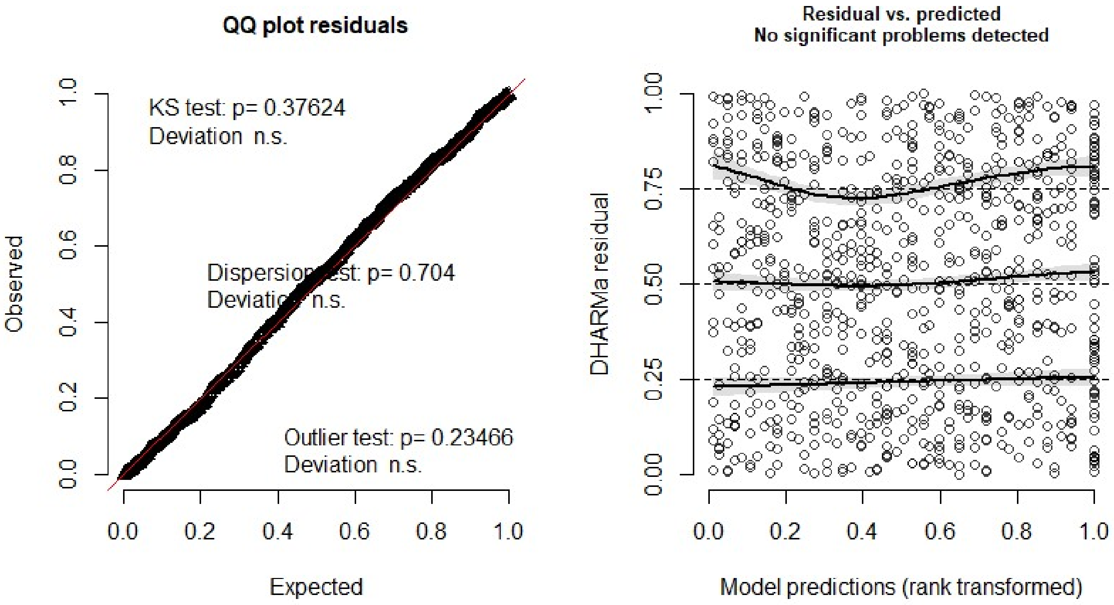
[1] Jemila S. Hamid, Wei Liang Huang & Dietrich von Rosen (2020) ”Graphical analysis of residuals in multivariate growth curve models and applications in the analysis of longitudinal data”.

Figure 9. ACF (left) and QQ-plot (right) of the random effects from the fitted model ($glmm3$).



with the $DHARMa$[2] package, which "uses a simulation-based approach to create readily interpretable scaled (quantile) residuals for fitted (generalized) linear mixed models". The DHARMa analysis (see Figure 10) does not show any sign of dispersion issues, nor outlier, nor any other type of problems on the residuals of the fitted model ($glmm3$).

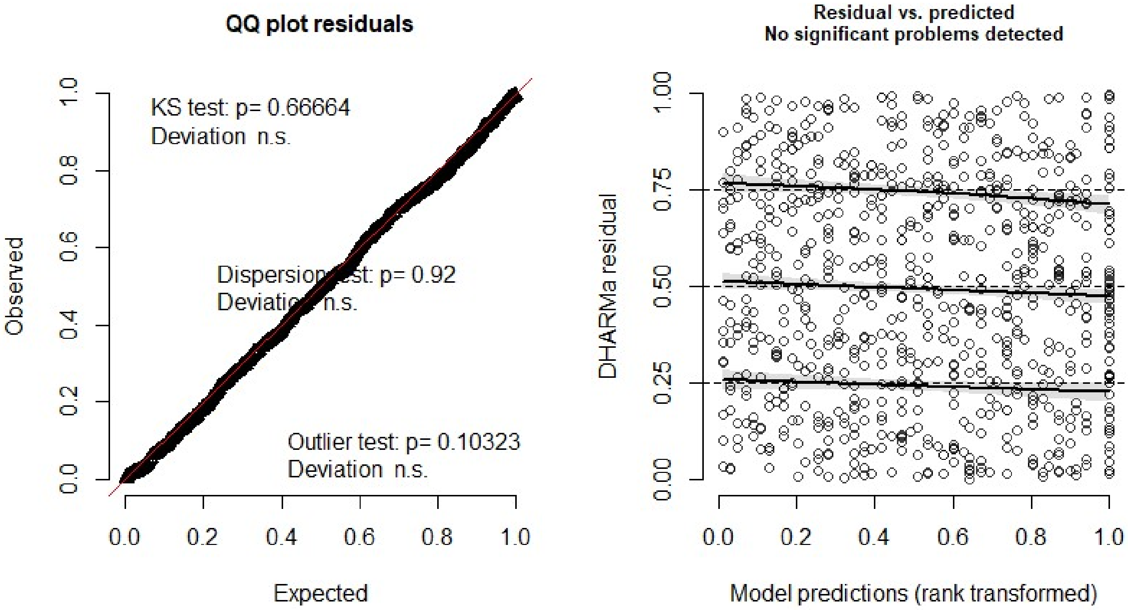
Figure 10. DHARMa residual analysis of the fitted model ($glmm3$). Left: QQ-plot to detect overall deviations from the expected distribution, with added tests for correct distribution (KS test), dispersion, and outliers. Right: plot of the residuals against the predicted value.



In an attempt to improve the goodness of fit, a new model ($glmm3\_new$) is fitted without the patients with the largest five random effects (see Figure 8). The DHARMa analysis (see Figure 11) on the residuals of the new model shows arguably better results (the lines on the right panel are straighter), but does not exhibit a noteworthy improvement. Both models ($glmm3$ and $glmm3\_new$) are considered to be acceptable in terms of their fit to the data.

[2] DHARMa: residual diagnostics for hierarchical (multi-level/mixed) regression models.

Figure 11. DHARMa residual analysis of the new fitted model (*glmm3_new*). Left: QQ-plot to detect overall deviations from the expected distribution, with added tests for correct distribution (KS test), dispersion, and outliers. Right: plot of the residuals against the predicted value.



2.6 Inference & Interpretation

The p-values of the fixed effect coefficients (see Figure 7) of the fitted model (*glmm*3) show which terms are statistically significant. As expected, the time variable $MONTH$ is significant at the 1% level with a negative value, implying that the probability of presence of thought disorders decreases over time. On average, the effect of the observation month appears to be strong, with a standard error comparatively small. However, the GLMM is a subject-specific analysis, and the coefficients do not necessarily apply to the population. The random intercept standard deviation is 2.313, three times larger than the observation month coefficient, meaning that there is considerable variation among individuals.

Another significant variable is the interaction term between $MONTH$ and $GENDER$, which indicates whether the evolution of thought disorders linearly depends upon the specific gender of the patient. Its positive estimated value suggests that the prevalence of thought disorders evolves less favorably for a 'male' individual (coded as 1) than for a 'female' individual (coded as 0). In fact, when the log(odds) of thought disorders for two consecutive months is calculated for a 'male' patient and a 'female' patient in the same age category, the log(odds) for the 'female' patient decreases 0.283 more than for the 'male' patient, which corresponds to the coefficient of the $MONTH \cdot GENDER$ term in eq.(2).

3 Conclusion

In the estimated parameters of the GLMM two terms are found statistically significant: the time variable $MONTH$, and the interaction term between $MONTH$ and $GENDER$. The latter indicates that the evolution differs for each gender group at the subject-specific level. More precisely, the evolution of 'male' patients is less favorable than the one of 'female' patients. Lastly, the interaction term between $MONTH$ and AGE is not statistically significant, meaning that no weighty difference is found among the evolution of 'young' and 'old' patients.