Longitudinal Analysis of Serum Bilirubin Trajectories in Primary Biliary Cirrhosis

Yixin Zheng

2025-12-14

1 Abstract

This study aimed to describe how log-transformed serum bilirubin levels change over time in patients with primary biliary cirrhosis and to find the best way to model repeated measurements from the same patient. Analyzing data from the randomized D-penicillamine trial, we found that a random-intercept model with exponential residual correlation and a nugget effect best fit the data. A linear time trend was enough to describe the average change in bilirubin levels.

2 Introduction

PBC is a chronic liver disease where liver function gradually deteriorates. Serum bilirubin is an important marker for disease severity, and tracking the changes in serum bilirubin level helps to illustrate the natural history of the disease. Repeated measurements carry more information than a single first measurement alone. Biomarker data collected over time involves repeated measurements on the same individual that are interrelated and may differ significantly between individuals. Ignoring such patterns can lead to misleading results. Linear mixed-effects models take these relationships and differences into account in the analyses of such data. This analysis is based on data from a randomized D-penicillamine trial in PBC patients (pbcseq dataset, R survival package), which contains repeated laboratory measurements over time and is hence ideal for illustrating how to work with data collected over time. The key scientific and statistical question being addressed in this analysis is: What is the average longitudinal profile of log-transformed serum bilirubin in patients with PBC, and what dependence structure within the subjects best describes the repeated measures? To answer this question, we concentrate on two formal sub-questions: Mean structure: Is a linear time trend sufficient to model the marginal mean trajectory of log(serum bilirubin), or is there data-based support for a more general cubic time trend? Dependence structure: What is the best dependence structure for the within-subject correlation in the bilirubin measurements after accounting for subject-specific heterogeneity, which ranges from independent errors to random-effects and continuous-time residual correlation models?

3 Methods

3.1 Outcome

The primary longitudinal outcome was serum bilirubin, analyzed on the log scale, log(bili), to reduce skewness and improve consistency. We recorded follow-up time in days from enrollment and converted it to years (day/365.25) for interpretability. Follow-up time was centered at its sample mean,

$$t^* = t - \bar{t}$$

, to improve numerical stability and make the model intercepts easier to understand.

3.2 Data Cleaning

The `pbcseq` dataset includes repeated lab measurements from the randomized D-penicillamine trial in patients with primary biliary cirrhosis. We restrict the data to randomized subjects with a recorded treatment assignment and to visits at or after baseline ($day \geq 0$). To emphasize the repeated-measures structure, we kept subjects who had at least 3 post-baseline bilirubin measurements. After filtering, we removed any unused factor levels.

3.3 Exploratory Data Analysis

For exploratory data analysis, we

(i) checked marginal distributions and summary statistics as basic sanity checks. This included the number of subjects and observations, the distribution of repeated measurements per subject, the range of follow-up times, the marginal distribution of log-transformed serum bilirubin, and treatment levels.

(ii) explored the marginal mean structure by plotting log(bilirubin) against follow-up time for all observations. We also plot individual trajectories for a random subset of subjects to aid visualization.

(iii) assessed within-subject dependence by fitting a marginal linear model with independent errors. We then calculated the lag-1 correlation of within-subject residuals and examined the residual autocorrelation function (ACF) to assess autocorrelation across multiple lags.

3.4 Models

3.4.1 Mean Structure

Let Y_{ij} denote log(bilirubin) for subject i measured at follow-up time t_{ij} (years since baseline). We center time at the sample mean,

$$t_{ij}^* = t_{ij} - \bar{t}$$

.

We consider two candidate marginal mean models: (1) linear: $E(Y_{ij}) = \beta_0 + \beta_1 t_{ij}^*$; (2) cubic: $E(Y_{ij}) = \beta_0 + \beta_1 t_{ij}^* + \beta_2 (t_{ij}^*)^2 + \beta_3 (t_{ij}^*)^3$.

3.4.2 Dependence Structure

Under the cubic mean model, we compare dependence structures fit by maximum likelihood (ML) to enable AIC comparisons: (1) independent errors (GLS); (2) random intercept (RI); (3) random intercept + random slope in time (RI/RS); (4) RI + continuous-time AR(1) residual correlation (CAR(1)); (5) RI + exponential residual correlation; (6) RI + exponential residual correlation with a nugget effect.

3.5 Model Selection

Dependence structures were compared under the cubic mean model using maximum likelihood estimation to enable comparison via Akaike's Information Criterion (AIC). The model with the smallest AIC was selected as the preferred dependence structure. Holding this dependence structure fixed, the linear and cubic mean models were compared using a likelihood ratio test, with both models fit using maximum likelihood.

The null and alternative hypotheses are

$$H_0 : \ \beta_2 = \beta_3 = 0 \quad \text{(linear mean)}, H_A : \ (\beta_2, \beta_3) \neq (0, 0) \quad \text{(cubic mean)}$$

The likelihood ratio test statistic is

$$\Lambda = 2 \left\{ \ell(\widehat{\theta}_{\text{cubic}}) - \ell(\widehat{\theta}_{\text{linear}}) \right\} \sim \chi^2_2 \quad \text{under } H_0.$$

3.6 Diagnostics

Model diagnostics were conducted for the selected final model. These included examination of normalized residuals versus fitted values, normal Q–Q plots of residuals, residual autocorrelation functions, and residuals plotted against follow-up time. These diagnostics were used to assess mean specification, distributional assumptions, and the presence of remaining within-subject dependence.

4 Results

4.1 Exploratory Data Analysis

4.1.1 Marginal distributions / sanity check

After filtering, the analytic dataset includes 128 subjects and 935 observations. Each subject has at least 3 repeated measurements, with a median of 7 and a maximum of 16 visits. The mean of the centered follow-up time is approximately 0, confirming correct centering.

Follow-up time ranges from 0 to 13.90 years after baseline, and all visits are at or after baseline (day 0 or later). The log-transformed serum bilirubin values range from -1.61 to

3.63, with a median of 0.26 and a mean of 0.58. Figure 1 shows the marginal distribution of log(bilirubin).

After keeping only subjects with at least three visits, the dataset includes only the D-penicillamine treatment arm. As a result, treatment effects are not included in later analyses.

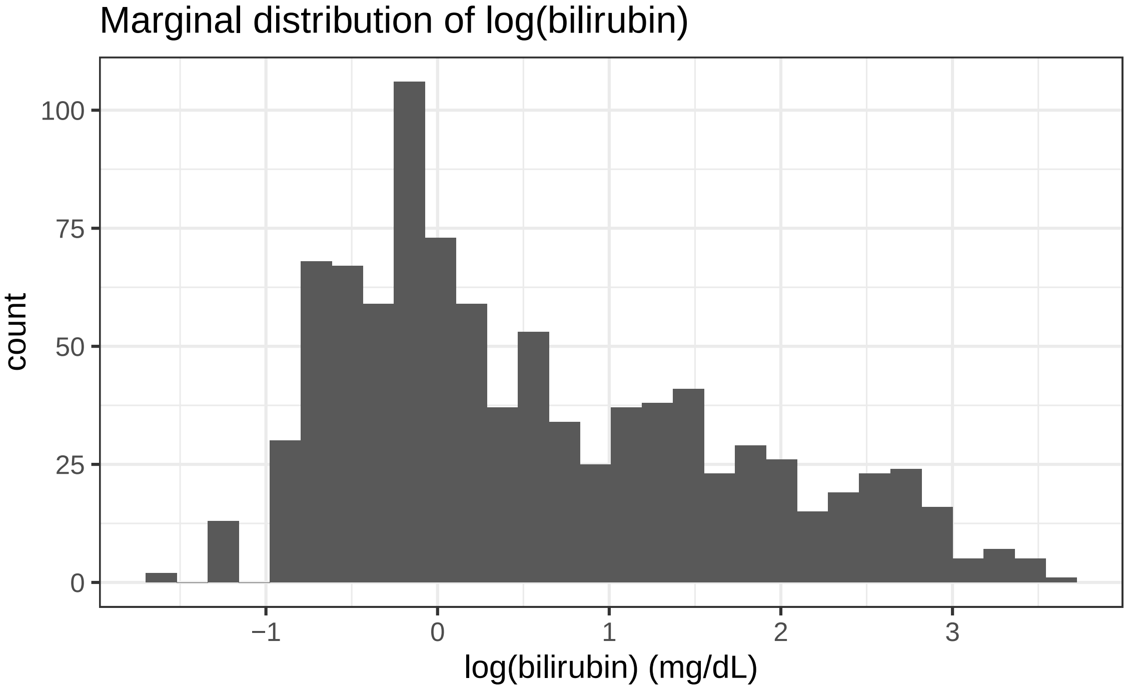

Figure 1: Exploratory Plot: Marginal Distribution

Figure 1 shows the marginal distribution of log(bilirubin). The marginal distribution of log(bilirubin) is right-skewed, with most observations concentrated between approximately -0.5 and 1.5. A smaller number of observations extend into higher values, up to about 3.5

4.1.2 Mean Structure (population trend + individual trajectories)

Figure 2 shows log(bilirubin) over follow-up time for all observations. It also includes individual trajectories for a random sample of subjects. Over the follow-up period, log(bilirubin) values vary widely across subjects. Each person starts at a different baseline, and their levels change in different ways over time. Some people's values go up, while others stay about the

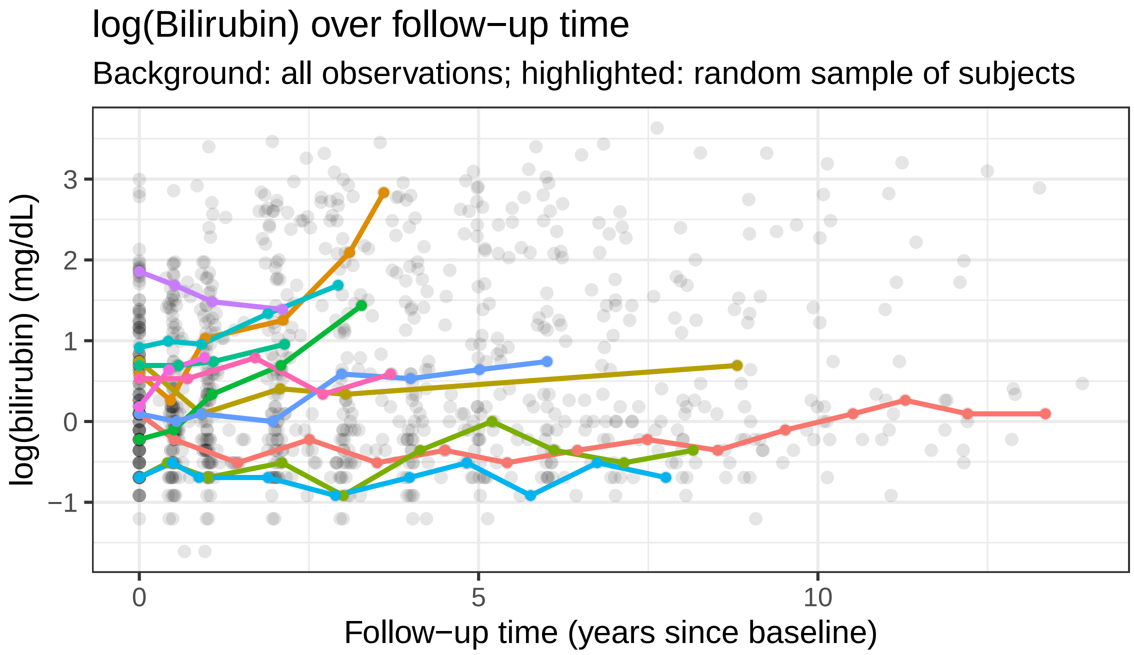

Figure 2: Exploratory Plot: Mean Structure

same or change in other ways, showing that each subject has a unique pattern.

4.1.3 Dependence Structure (quantitative check via lag-1 residual correlation)

Table 1: Lag-1 residual correlation from marginal linear model (EDA)

cor_lag1	median_dt	n_pairs
0.926	0.977	807

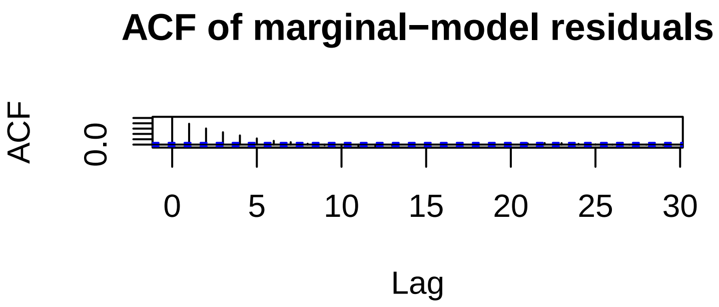

Figure 3: Exploratory Plot: ACF

A marginal linear model with independent errors was used to assess within-subject depen-

Table 2: Dependence structure comparison under cubic mean (ML)

Dependence	logLik	AIC
6. RI + Exp + nugget	-656.16	1328.32
3. RI + RS (t_star)	-675.55	1367.10
5. RI + Exp	-683.94	1381.88
4. RI + CAR(1)	-683.94	1381.88
2. RI	-879.09	1770.17
1. Independent (gls)	-1420.29	2850.59

```
## [1] "6. RI + Exp + nugget"
```

In the cubic mean model, six possible dependence structures were fitted using maximum likelihood and compared using AIC (see Table 2). The model with the lowest AIC was: Random intercept + exponential residual correlation with nugget effect (RI + Exp + nugget) (AIC = 1328.32, logLik = –656.16) All other candidate structures had higher AIC values, including the random intercept + random slope model and the random intercept + CAR(1) model.

Let $Y_{ki} = \log(\text{bili}_{ki})$ be serum bilirubin for subject k at visit i, with centered time

$$t^*_{ki} = t_{ki} - \bar{t}.$$

The fitted model is

$$Y_{ki} = \beta_0 + \beta_1 t^*_{ki} + \beta_2 (t^*_{ki})^2 + \beta_3 (t^*_{ki})^3 + b_{0k} + \varepsilon_{ki}, \qquad b_{0k} \sim N(0, \tau_0^2).$$

The selected dependence structure is

$$\mathrm{Corr}(\varepsilon_{ki}, \varepsilon_{kj}) = \exp\Big(-\phi|t^*_{ki} - t^*_{kj}|\Big) + \delta_{ij}\sigma^2_{\mathrm{indep}},$$

with $\delta_{ij} = 1$ if $i = j$ and 0 otherwise.

4.3 Mean Structure Comparison

```
##                   Model df      AIC      BIC    logLik   Test  L.Ratio p-value
## fit_linear_best      1  6 1328.206 1357.249 -658.1029
## fit_cubic_best       2  8 1328.320 1367.044 -656.1599 1 vs 2 3.886007  0.1433
```

A likelihood ratio test was used with the chosen dependence structure (RI + Exp + nugget) to compare the linear and cubic mean models.

The likelihood ratio test comparing the linear and cubic mean models gave the following results: LR = 3.89, df = 2, p = 0.143. The AIC values were 1328.21 for the linear model and 1328.32 for the cubic model.

4.4 Final Model

Based on the likelihood ratio test, the linear mean model was retained. The final fitted model is

$$Y_{ki} = \beta_0 + \beta_1 t^*_{ki} + b_{0k} + \varepsilon_{ki}, \qquad b_{0k} \sim N(0, \tau_0^2),$$

with exponential residual correlation and a nugget effect.

The fitted marginal mean is

$$\widehat{E}(Y_{ki}) = \widehat{\beta}_0 + \widehat{\beta}_1 t^*_{ki} = 0.7918 + 0.1095\, t^*_{ki}$$

The estimated fixed effects show a positive linear association between centered follow-up time and log-transformed serum bilirubin.

4.4.1 Fixed Effect estimates

Table 3: Estimated fixed effects (linear model)

Term	Estimate	Std_Error
(Intercept)	0.7918	0.0954
t_star	0.1095	0.0126

The final model chosen was a linear mean model with random intercepts and an exponential residual correlation + nugget effect.

Table 3 shows the estimated fixed effects from this model.

4.4.2 Fixed effect trajectories

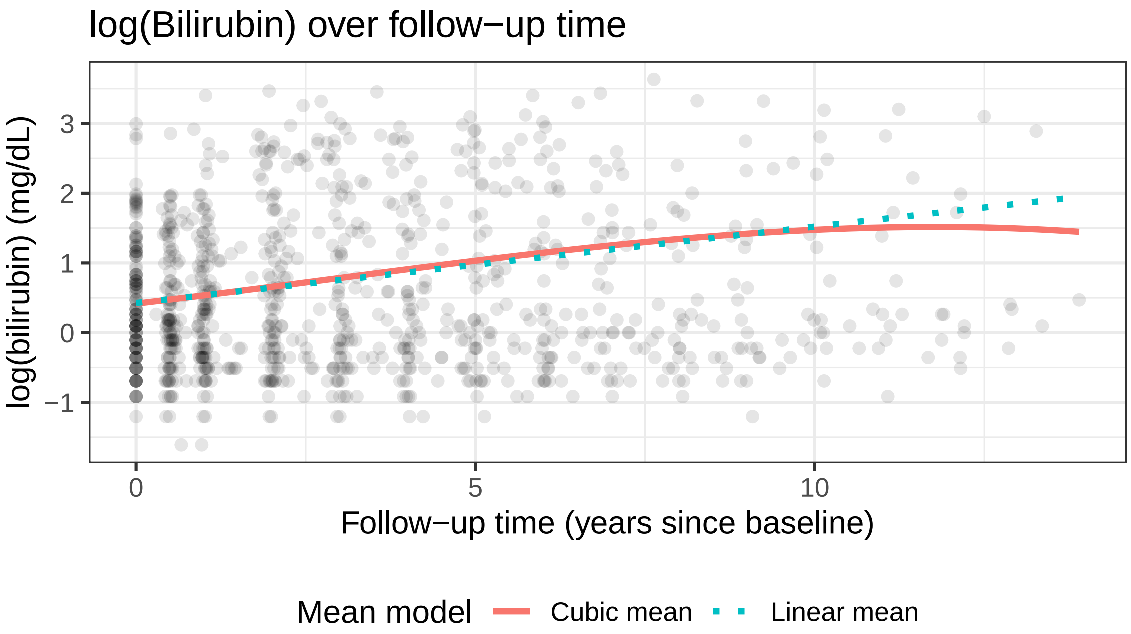

Figure 4: Fitted Marginal Mean Curves

Figure 4 shows the fitted marginal mean curves for both the linear and cubic models over the follow-up period. Both the linear and cubic fitted marginal mean curves show that log(bilirubin) increases over time. The cubic curve bends slightly at later follow-up times compared to the linear fit, but the two curves are similar for most of the observed time range

4.5 Diagnostics

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## -3.9738 -0.6219 -0.1114 -0.0121  0.5054  4.8559
```

The summary statistics for the normalized residuals showed a mean close to 0, with values ranging from –3.974 to 4.856.

Figure 5 shows the diagnostic plots for the selected model:

1. Residuals vs fitted values show mild heteroskedasticity.

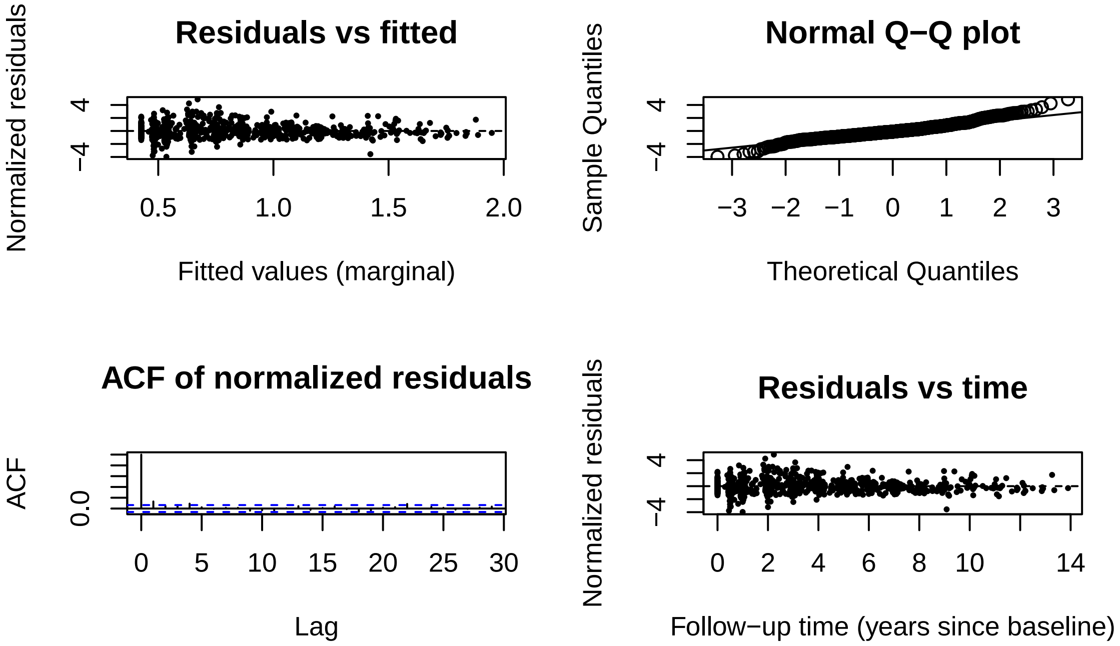

Figure 5: Diagnostic Plots

2. Normal Q–Q plot shows that the tails deviate from normality.
3. ACF of normalized residuals shows much less autocorrelation compared to the marginal model.
4. Residuals vs time show no strong remaining time trends.

5 Discussion

We explored the longitudinal profiles of log-transformed serum bilirubin in patients with primary biliary cirrhosis, employing mixed-effects models, with varying marginal mean and within-subject dependence structures, in order to determine the best-fitting models. Preliminary exploratory analyses indicated significant heterogeneity in baseline bilirubin levels along with serial correlation in repeated measurements within subjects. Strong lag-1 residual correlation and a slowly decaying autocorrelation under a marginal linear model suggested the need for models containing explicit within-subject dependence rather than an independent-error assumptions.

Comparing models with a cubic marginal mean, we found that accounting for both subject-specific random intercepts and continuous-time residual correlation represented an improvement in model fit compared to less complex alternatives. The random-intercept model with exponential residual correlation and a nugget effect had the lowest AIC of the three, suggesting that both ongoing within-subject correlation and variability at the level of measurement are critical components in the data. Models with only independent errors or random intercepts did not fully capture the dependence in the longitudinal bilirubin measurements.

When the marginal mean structures were compared using the selected dependence model, the likelihood ratio test comparing the cubic time trend to the linear time trend was not statistically significant, and the linear marginal mean was therefore kept The cubic model, although less constrained, revealed a curve with an upward trend for most follow-up times, similar to the linear model. This implies that, for these data, a simple linear time effect may be sufficient to summarize the average log(bilirubin) trajectory when within-subject dependence is modeled appropriately.

The final model selected captures the following three key features of the data: Differences in baseline bilirubin levels across subjects, represented by a random intercept; serial correlation in repeated measurements over time, expressed as exponential correlation; and extra observation-level variability, modeled as a nugget effect. Diagnostic plots of this model indicate better residual behavior than simpler models, with less autocorrelation and no strong systematic pattern over time, supporting the effectiveness of this model.

There are some limitations to consider. After the filtration, only patients with D-penicillamine treatment were left; hence, it was not possible to study the effects of treatment or the interaction between treatment and time. We also used time as the only predictor in our model; incorporation of disease severity or other clinical information could further enhance the precision of the estimated trends.

In a nutshell, the current study shows that careful exploratory analysis and thoughtful model comparison help choose the right ways of describing both average trends and within-person patterns in longitudinal data. The findings emphasize the need to consider serial correlation and patient differences while modeling repeated biomarker measurements in chronic diseases like primary biliary cirrhosis.