Longitudinal Analysis on Various Factors for Human Immune System

Younghwan Brian Cho

Abstract

This report describes a parametric model (Linear Mixed Model) for the longitudinal data. In our case, we wish to analyze effects of several variables on human immune system (CD4+ cells) after the infection of HIV. The essential work is to identify proper mean function, the form of random effects, the covariance of the random effects and the covariance of the random errors. Since AIDS is still incurable disease, the final results suggest that the time is the most important and significant variable. In other words, the number of CD4+ cells gradually decreases and the decreasing speed accelerates as time goes by. Other predictors including smoking (sig), recreational drugs (drug) and the minus score of center for epidemiological depression (cesd) will also have a positive effects to patients. Finally, the model can also be applied to predict the remaining survival time of the individual patients.

1. Introduction

1.1 HIV and CD4 Cells

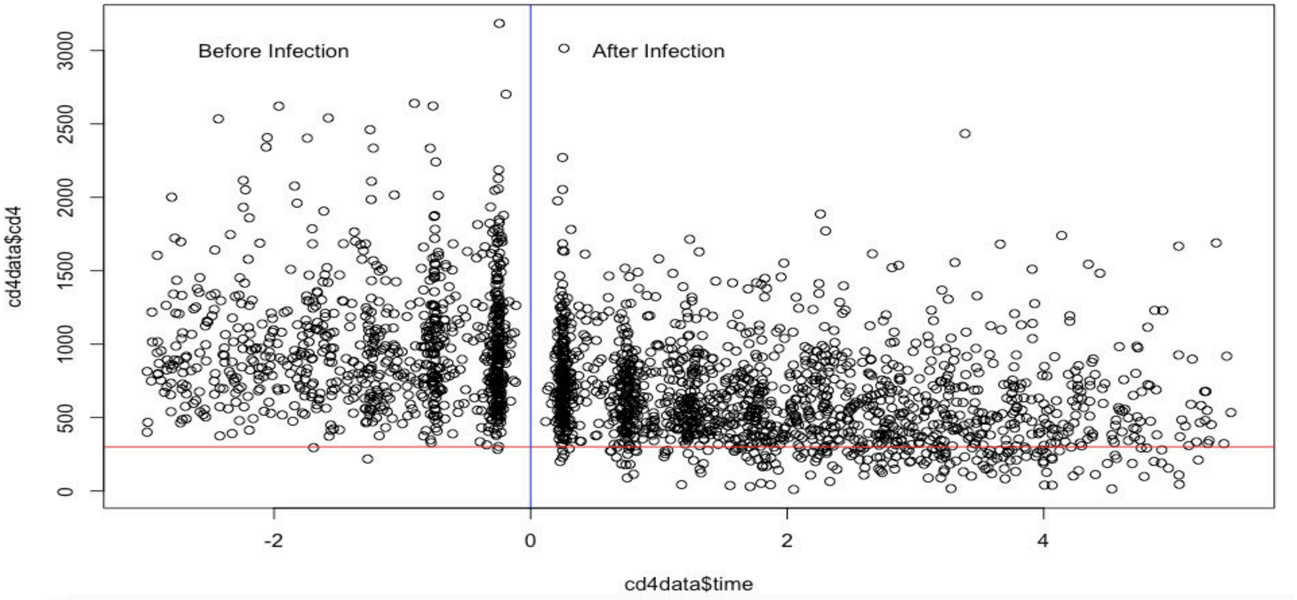
HIV destroys T-lymphocytes called CD4 cells, which play a vital role in human immune function. Disease progression can be assessed by measuring the number or percent of CD4 cells remaining in our body, which on average decreases throughout the disease incubation period. In addition, CD4 cell levels may be related to cofactors[1] such as age, smoking or new therapies. Generally speaking, when the number of CD4+ cells is less than 300 per unit (the red line in the graph below), which means high risk of developing serious illness, such as AIDS, or even death.

1.2 Data Background

The motivating data are from the Multicenter AIDS Cohort Study or "MACS," which has followed nearly 5,000 gay and bisexual men from Baltimore, Pittsburgh, Chicago, and Los Angeles since 1984. 2,376 measurements of CD4 cell numbers plotted against time since seroconversion (t = 0) for 369 seroconverts. Seroconversion is the period during which HIV antibodies develop and become detectable. Also, personal habits or traits of the subjects including smoking habit, drug usage and cesd have been recorded in order to determine if these traits affect positively or negatively on the

[1] Zeger and Diggle (1994) Semiparametric Models for Longitudinal Data with Application to CD4 Cell Numbers in HIV Seroconverters

speed of CD4 cell count drop. More details of the study design and methods are reported by Kaslow[2].



2. Model Formulation

2.1 Linear Mixed Model

The parametric model we choose to fit the CD4+ data set with related covariates is the Linear Mixed Model (LMM), which is based on the models setting[3] given by Breslow, Clayton 1993. The main R package we applied for generalized linear mixed model is "nlme[4]".

$Y_{ij} = the\ response\ of\ j^{th}\ case\ of\ the\ i^{th}\ cluster.$

$i = 1,\ \dots,\ m\ ,\ \ j = 1,\ \dots,\ n_i\ .$

$m = the\ number\ of\ the\ cluster\ (patients).$

$n_i = the\ size\ of\ the\ cluster\ (records\ of\ i^{th}\ patient).$

$x_{ij} = the\ covariate\ vector\ of j^{th}\ case\ of\ the\ i^{th}\ cluster$ for fixed effects.

$\beta = fixed\ effects\ parameter\ \in R^p.$

$u_{ij} = covariate\ vector\ of j^{th}\ case\ of\ the\ i^{th}\ cluster\ for\ random\ effects$

$\gamma_i = radom\ effects\ parameter \in R^q.$

[2] Kaslow et al. (1987) Competencies in Professional Psychology.

[3] N. E. Breslow and D. G. Clayton (1993). Approximate Inference in Generalized Linear Mixed Models

[4] José Pinheiro, Douglas Bates, et al, Linear and Nonlinear Mixed Effects Models

The LMM is defined as below with fixed effects, random effects and random error:

$$Y_{ij} = x_{ij}^t\beta + u_{ij}^t\gamma_i + \epsilon_{ij} = fixed\ effects + random\ effects + random\ error\ ,$$

where $\gamma_i \sim N_q(0, G)$, $G \in R^{q\times q}$, $G = the\ covariance\ matrix\ of\ the\ random\ effects\ \ \gamma_i$,

$\epsilon_i = \begin{pmatrix} \epsilon_{i1} \\ \vdots \\ \epsilon_{in_i} \end{pmatrix} \sim N_{ni}(0, \Sigma_i)$, $\Sigma_i \in R^{n_i\times n_i}$, Σ_i= the covariance matrix of the random errorϵ_i ,

$\gamma_1 , \dots , \gamma_m , \ \epsilon_1, \ \dots , \ \epsilon_m : independent$ **.**

2.2 Response Transformation

First, we fit a simple linear model to check the performance of the response. If the relation is not linear enough, then we need to apply some other transformation methods.

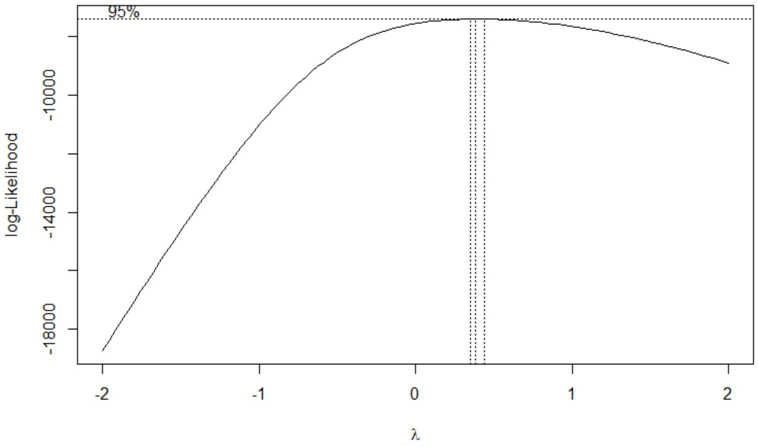
The linear model we fit is shown:

$$CD4Cell_{ij} = \beta_0 + \beta_1\mathrm{time}_{i,j} + \beta_2\mathrm{time}_{i,j}^2 + \beta_3\mathrm{time}_{i,j}^3 + \beta_4\mathrm{cig}_{i,j} + \beta_5\mathrm{drug}_{i,j} + \beta_6\mathrm{cesd}_{i,j} + \beta_7\mathrm{cesd}_{i,j} \times \mathrm{time}_{i,j} + \varepsilon_{i,j}$$

Based on the Box-Cox Transformation[5]

$$y_i^{(\lambda)} = \begin{cases} \dfrac{y_i^\lambda - 1}{\lambda}, if\ \lambda \neq 0\ , \\ In(y_i)\ , if\ \lambda = 0. \end{cases}$$

The log-Likelihood function with different lambda are shown below:

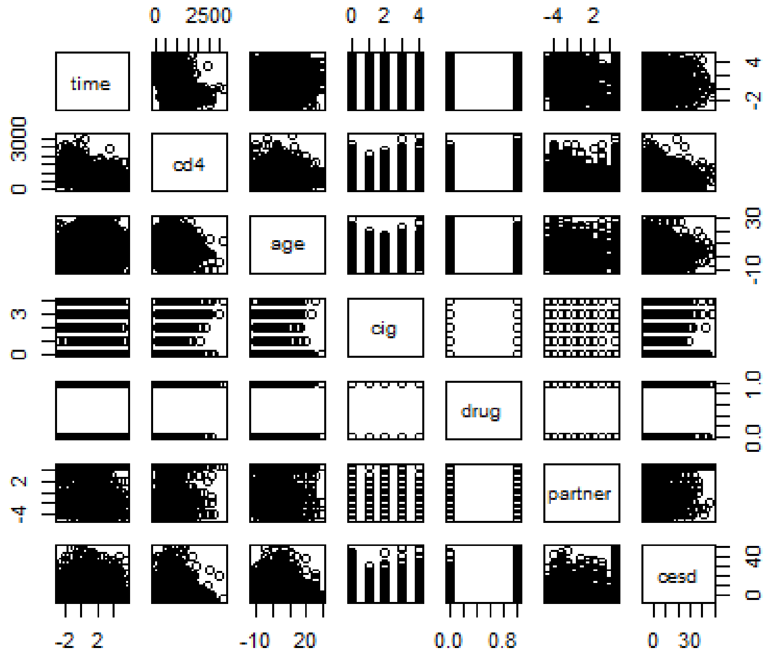


Based on the graph, estimated parameter: $\hat{\lambda} = \frac{1}{2}$.

[5] G. E. P. Box and D. R. Cox (1964). An Analysis of Transformations

2.3 Correlation between Predictors

Before we decide mean function, we first check if there is significant correlation between predictors. Below is the scatter plot matrix between predictors.



From this we conclude there is no outstanding linear relationship between predictors.

2.4 Mean Function

Applying the transformed data set with power = 0.5, we refit several linear models to detect the suitable mean function for the LMM. These 2 candidate models have been considered and shown below:

cd4lm1: $\sqrt{CD4Cell_{ij}} = \beta_0 + \beta_1 \mathrm{time}_{i,j} + \beta_2 \mathrm{cig}_{i,j} + \beta_3 \mathrm{drug}_{i,j} + \beta_4 \mathrm{cesd}_{i,j} + \varepsilon_{i,j}$

cd4lm2: $\sqrt{CD4Cell_{ij}} = \beta_0 + \beta_1 \mathrm{time}_{i,j} + \beta_2 \mathrm{time}_{i,j}^2 + \beta_3 \mathrm{time}_{i,j}^3 + \beta_4 \mathrm{cig}_{i,j}$

$$+\beta_5 \mathrm{drug}_{i,j} + \beta_6 \mathrm{cesd}_{i,j} + \beta_7 \mathrm{time}_{i,j} \times \mathrm{cig}_{i,j} + \varepsilon_{i,j}$$

We apply F-test to compare these 2 models. The null hypothesis of the F-test[6] is:

H_0: the reduced model (cd4lm1) is as good as the full model (cd4lm2).

$H_1: the\ full\ model\ (cd4lm2) is\ better\ than\ the\ reduced\ model\ (cd4lm1).$

[6] Lomax, Richard G. (2007). Statistical Concepts: A Second Course

F Test Statistics:

$$F = \frac{\dfrac{RSS_{reduce} - RSS_{full}}{df_{full} - df_{reduce}}}{\dfrac{RSS_{full}}{n - df_{full}}} \sim F_{df_{full} - df_{reduce}, n - df_{full}}$$

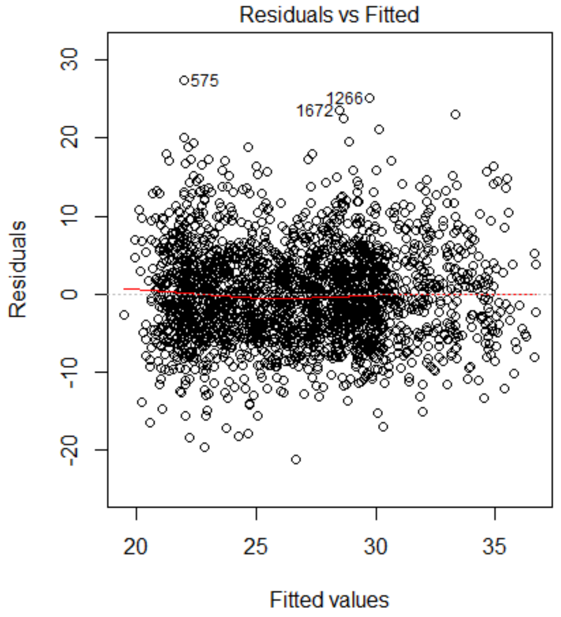
The test result is shown in the able below:

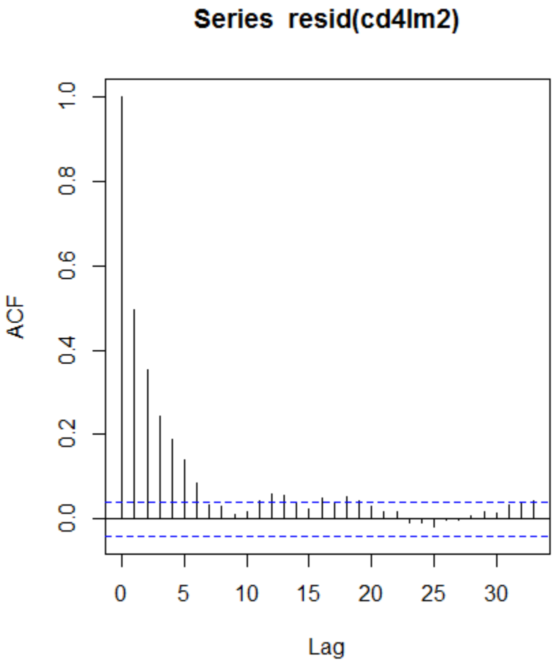
	df	RSS	$\boldsymbol{df_{full} - df_{reduce}}$	Diff RSS	F	Pr(>F)
cd4lm1	2371	88841				
cd4lm2	2368	85482	3	3358.1	31.008	< 2.2e-16

Since the P-value of the F-test is much smaller than the significance level ($\alpha = 0.05$), we have significant statistical evidence to reject the null hypothesis. Considering the fact that more covariates (information) have been used in the full model, we could conclude the full model (cd4lm2) is better than the reduced model (cd4lm1).

2.6 Covariance Structure of Random Errors

We check the residual plots and autocorrelation plot to detect the correlation structure for the random error.





The residual plot shows no pattern and constant variance. But the ACF plot of residuals show a clear autocorrelation, which is expected. Generally, each patients have 4 or 5 records. Each record must be highly auto-correlated to previous step, since progression of disease is gradual and affected by past condition. So, an AR1 covariance structure is applied to the error term for each cluster\patient.

$$\varepsilon_i \sim N_{n_i}(0, \Sigma_i) \ ,$$

$$\Sigma_i = \sigma^2 \begin{bmatrix} 1 & \rho^{T_{1,2}} & \rho^{T_{1,3}} & \cdots & \rho^{T_{1,n_i}} \\ \rho^{T_{2,1}} & 1 & \rho^{T_{2,3}} & \cdots & \rho^{T_{2,n_i}} \\ \rho^{T_{3,1}} & \rho^{T_{3,2}} & 1 & \cdots & \rho^{T_{3,n_i}} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho^{T_{n_i,1}} & \rho^{T_{n_i,2}} & \rho^{T_{n_i,3}} & \cdots & 1 \end{bmatrix}_{n_i \times n_i} \ ,$$

where ρ is the autocorrelation for the random error for the records of the same patients,

$T_{i,j}\ is\ the\ (i,j)\ element\ of\ matrix\ T$,

$$T = \begin{bmatrix} 0 & |\mathrm{time}_{i,1} - \mathrm{time}_{i,2}| & \cdots & |\mathrm{time}_{i,1} - \mathrm{time}_{i,n_i}| \\ |\mathrm{time}_{i,1} - \mathrm{time}_{i,2}| & 0 & \cdots & |\mathrm{time}_{i,2} - \mathrm{time}_{i,n_i}| \\ |\mathrm{time}_{i,1} - \mathrm{time}_{i,3}| & |\mathrm{time}_{i,2} - \mathrm{time}_{i,3}| & \cdots & |\mathrm{time}_{i,3} - \mathrm{time}_{i,n_i}| \\ \vdots & \vdots & \vdots & \vdots \\ |\mathrm{time}_{i,1} - \mathrm{time}_{i,n_i}| & |\mathrm{time}_{i,2} - \mathrm{time}_{i,n_i}| & \cdots & 0 \end{bmatrix}_{n_i \times n_i} \ .$$

2.7 Random Effects

After specifying the mean function and the covariance structure of the random error, we apply the LMM with random effects. The next step is to specify the suitable form of random effects. We compared 2 potential candidate random effects (the mean function and setting for random error are the same), the first is only the individual intercept form, and the other is the individual simple linear form:

cd4lmm.1.I*:* $\sqrt{CD4Cell_{ij}} = \beta_0 + \beta_1 \mathrm{time}_{i,j} + \beta_2 \mathrm{time}_{i,j}^2 + \beta_3 \mathrm{time}_{i,j}^3 + \beta_4 \mathrm{cig}_{i,j}$

$$+ \beta_5 \mathrm{drug}_{i,j} + \beta_6 \mathrm{cesd}_{i,j} + \beta_7 \mathrm{time}_{i,j} \times \mathrm{cig}_{i,j} + \gamma_{0i} + \varepsilon_{i,j} \ ,$$

$$\gamma_{0i} \sim N\left(0, \sigma_{\gamma_0}^2\right) .$$

cd4lmm.un.ar1: $\sqrt{CD4Cell_{ij}} = \beta_0 + \beta_1 \mathrm{time}_{i,j} + \beta_2 \mathrm{time}_{i,j}^2 + \beta_3 \mathrm{time}_{i,j}^3 + \beta_4 \mathrm{cig}_{i,j}$

$$+\beta_5 \mathrm{drug}_{i,j} + \beta_6 \mathrm{cesd}_{i,j} + \beta_7 \mathrm{time}_{i,j} \times \mathrm{cig}_{i,j} + \gamma_{0i} + \gamma_{1i} \times \mathrm{time}_{i,j} + \varepsilon_{i,j} \ ,$$

$$\gamma_i = \begin{bmatrix} \gamma_{0i} \\ \gamma_{1i} \end{bmatrix} \sim N_q\left(0, G_{q \times q}\right) \ , \ q = 2 \ , G = \begin{bmatrix} \sigma_1^2 & r\sigma_1\sigma_2 \\ r\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix} .$$

In order to compare these 2 models, we use the likelihood ratio test[7] (LRT).

$$H_0{:}\ model\ 0\ is\ as\ good\ as\ model\ 1. \quad \text{Vs.} \quad H_1{:}\ model\ 0\ is\ not\ as\ good\ as\ model\ 1.$$

[7] Casella, George; Berger, Roger L. (2001). *Statistical Inference* (Second ed.).

$$LR = -2log\left[\frac{L(\hat{\beta}_0)}{L(\hat{\beta})}\right] = -2\left[L(\hat{\beta}_0) - L(\hat{\beta})\right] \sim X_{df}^2$$

Model cd4lmm.1.I is nested inside the model cd4lmm.un.ar1. The result of the LRT are shown in the table below.

	Model	df	logLikelihood	Test	Loglike Ratio	p-value
cd4lmm.1.I	0	11	-7313.362			
cd4lmm.un.ar1	1	13	-7309.448	0 vs 1	7.828	0.02

Because the p-values (0.02) from the table above is smaller than $\alpha = 0.05$, we conclude there is significant evidence to reject the null hypothesis, which means these 2 models are different in goodness of fit. Considering the extra information given by the individual slope of the random effects, we conclude the model (cd4lmm.un.ar1) with the individual simple linear random effects is much better.

2.8 Covariance Structure of Random Effects

After specifying the suitable for the random effects (individual simple linear), we need to improve the covariance structure for the random effects. Three potential covariance structures have been compared, and these are defined as below.

$$\sqrt{CD4Cell_{ij}} = \beta_0 + \beta_1 \text{time}_{i,j} + \beta_2 \text{time}_{i,j}^2 + \beta_3 \text{time}_{i,j}^3 + \beta_4 \text{cig}_{i,j}$$

$$+\beta_5 \text{drug}_{i,j} + \beta_6 \text{cesd}_{i,j} + \beta_7 \text{time}_{i,j} \times \text{cig}_{i,j} + \gamma_{0i} + \gamma_{1i} \times \text{time}_{i,j} + \varepsilon_{i,j} \ \ ,$$

where $\gamma_i = \begin{bmatrix} \gamma_{0i} \\ \gamma_{1i} \end{bmatrix} \sim N_q(0, G_{q \times q})$, $q = 2$, with 3 different G:

cd4lmm.un.ar1: $G = \begin{bmatrix} \sigma_1^2 & r\sigma_1\sigma_2 \\ r\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}$

cd4lmm.diag.ar1: $G = \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix}$

cd4lmm.inde.ar1: $G = \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_1^2 \end{bmatrix}$

Since these 3 covariance structure are nested, we use LRT to compare the same LMM with different covariance structures for random effects. The test results are shown in the table below.

	Model	df	LogLik	Test	Likli Ratio	P-values
cd4lmm.un.ar1	1	13	-7309.448			
cd4lmm.diag.ar1	2	12	-7309.649	1 vs 2	0.4024	0.5258
cd4lmm.inde.ar1	3	11	-7321.639	2 vs 3	23.979	<0.0001

From the table above, we can conclude there is no significant difference between Model1 and Model2, however Model2 is significantly different from Model3. Considering Model1 is unstructured form for covariance, and considering the penalty for estimating more parameters than Model2, we conclude the Model2 (cd4lmm.diag.ar1) with a diagonal matrix form for the covariance structure of the random effects is the best among these 3 candidates.

2.9 Final Model

After specifying the mean function, covariance for random error, random effects, covariance for random effects of the LMM, we are able to write down our final LMM (cd4lmm.diag.ar1):

$$\sqrt{CD4Cell_{ij}} = \beta_0 + \beta_1 \mathrm{time}_{i,j} + \beta_2 \mathrm{time}_{i,j}^2 + \beta_3 \mathrm{time}_{i,j}^3 + \beta_4 \mathrm{cig}_{i,j}$$

$$+\beta_5 \mathrm{drug}_{i,j} + \beta_6 \mathrm{cesd}_{i,j} + \beta_7 \mathrm{time}_{i,j} \times \mathrm{cig}_{i,j} + \gamma_{0i} + \gamma_{1i} \times \mathrm{time}_{i,j} + \varepsilon_{i,j}$$

$$\gamma_i = \begin{bmatrix} \gamma_{0i} \\ \gamma_{1i} \end{bmatrix} \sim N_q\left(0, G_{q\times q}\right) \ , \ q = 2 \ , G = \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix}, \varepsilon_i \sim N_{n_i}\left(0, \Sigma_i\right) ,$$

$$T = \begin{bmatrix} 0 & |\mathrm{time}_{i,1} - \mathrm{time}_{i,2}| & \cdots & |\mathrm{time}_{i,1} - \mathrm{time}_{i,n_i}| \\ |\mathrm{time}_{i,1} - \mathrm{time}_{i,2}| & 0 & \cdots & |\mathrm{time}_{i,2} - \mathrm{time}_{i,n_i}| \\ |\mathrm{time}_{i,1} - \mathrm{time}_{i,3}| & |\mathrm{time}_{i,2} - \mathrm{time}_{i,3}| & \cdots & |\mathrm{time}_{i,3} - \mathrm{time}_{i,n_i}| \\ \vdots & \vdots & \vdots & \vdots \\ |\mathrm{time}_{i,1} - \mathrm{time}_{i,n_i}| & |\mathrm{time}_{i,2} - \mathrm{time}_{i,n_i}| & \cdots & 0 \end{bmatrix}_{n_i \times n_i} \quad \Sigma_i = \sigma^2 \begin{bmatrix} 1 & \rho^{T_{1,2}} & \rho^{T_{1,3}} & \cdots & \rho^{T_{1,n_i}} \\ \rho^{T_{2,1}} & 1 & \rho^{T_{2,3}} & \cdots & \rho^{T_{2,n_i}} \\ \rho^{T_{3,1}} & \rho^{T_{3,2}} & 1 & \cdots & \rho^{T_{3,n_i}} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho^{T_{n_i,1}} & \rho^{T_{n_i,2}} & \rho^{T_{n_i,3}} & \cdots & 1 \end{bmatrix}_{n_i \times n_i} .$$

3. Model Diagnostic Checking

3.1 Model Selection by AIC

Although we have decided our final model by deciding each part of the LMM, and finally combine them together. We also conduct the AIC[8] to test if our Final Model is better than other compared models, based on the response after the Box Cox transformation.

$$\mathrm{AIC} = 2k - 2\ln(\hat{L})$$

The table below shows the AIC of different models:

Models	df	AIC
cd4lm1	6	15359.30
cd4lm2	9	15273.75
cd4lmm.1.I	11	14648.72
cd4lmm.un.ar1	13	14644.90

[8] Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle.

cd4lmm.diag.ar1	***12***	***14643.30***
cd4lmm.inde.ar1	11	14665.28

Since smaller AIC indicates better method, we can conclude our Final Model (cd4lmm.diag.ar1) is also consistent with the result from the AIC comparison.

3.2 Parameter Significance T Test for Mean Function

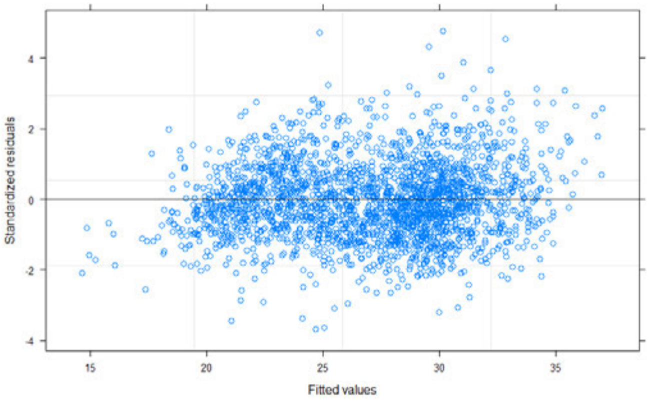
Parameter T test has been conducted to see significance of fixed effects of our final model cd4lmm.diag.ar1. Below is the test result.

All the parameters of fixed effects including time, time2, time3, cig, cesd and time:cig terms are found to be significant. Only drug term is found to be insignificant with p-value(0.1307) greater than 0.05.

```
                Value Std.Error   DF   t-value p-value
(Intercept) 28.527470 0.4105531 2000  69.48546  0.0000
time        -2.196960 0.1561324 2000 -14.07113  0.0000
time2       -0.440518 0.0549300 2000  -8.01963  0.0000
time3        0.118350 0.0141208 2000   8.38127  0.0000
cig          0.641313 0.1291044 2000   4.96740  0.0000
drug         0.473593 0.3132456 2000   1.51189  0.1307
cesd        -0.043725 0.0136695 2000  -3.19876  0.0014
time:cig    -0.186639 0.0586703 2000  -3.18114  0.0015
```

3.3 Residual Diagnostic

The residual plot of our Final Model below shows no certain pattern and shows constant variance. From this we conclude there is no major violation to assumptions for our model selection.



4. Conclusion

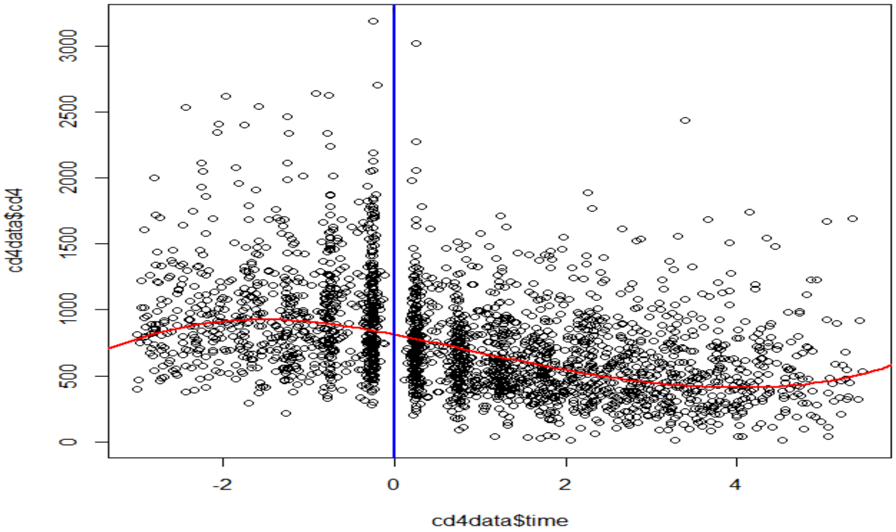
The parameter estimation for the mean function of our Final Model (cd4lmm.diag.ar1) based on the MLE method is shown in the table below.

Intercept	time	time2	time3	cig	drug	cesd	time:cig
$\widehat{\boldsymbol{\beta}}_{\mathbf{0}}$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	$\hat{\beta}_5$	$\hat{\beta}_6$	$\hat{\beta}_7$
28.527470	-2.196960	-0.440518	0.118350	0.641313	0.473593	-0.043725	-0.186639
P=0	0	0	0	0	0.13	0.0014	0.0015

4.1 time effect and trend

Based on our Final Model, the fitted values on time is drawn with the red curve below. From the graph, we can see the $\sqrt{CD4 + cells}$ will decrease $\left(\widehat{\beta_1} = -2.2\right)$ with time, and the decreasing speed will accelerate with quadratic term $\left(\widehat{\beta_2} = -0.44\right)$.

We suppose the tilted upward right tail is caused by the Survival Bias[9]: The logical error of concentrating on the people or things that made it past some selection process and overlooking those that did not, typically because of their lack of visibility.



4.2 other effects

The smoking gives a significant and positive effect ($\hat{\beta}_4 = 0.48$) on the patients. However, we need to note that since interaction of time and smoking ($\hat{\beta}_7 = -0.187$) is negative, smoking will have negative effect over the time. Suggestion for patient is that smoking is recommended at early stage of infection, but should be avoided over the period of disease progression.

[9] Elton; Gruber; Blake (1996). "Survivorship Bias and Mutual Fund Performance"

The estimated effects for recreational drug is ($\hat{\beta}_5 = 0.474$), which is positive. However, the parameter significance T test doesn't reject the null hypothesis of non-significance. Suggestion for patient is that, even though effect might insignificant recreational drug usage is recommended.

The minus score of center for epidemiological depression has a negative ($\hat{\beta}_6 = -0.0437$) effects on the patients, and this effects are significant. So, we would suggest patients to maintain positive mindset, otherwise progression of disease would proceed faster.