Linear Mixed Effects Models for CD4+ Cell Counts in Men with HIV

Anthony Anh Quoc Doan

May 4, 2018

Abstract

Human immunodeficiency virus or HIV are responsible for decline in CD4+ cell count. The investigation is set out to find the population rate of CD4+ cell count decline per milliliter of blood, to characterize the of individual rate of cell decline, and the factors that predict cell decline. Using exploratory data analysis and longitudinal tools, a linear mixed effects model with random intercept and random slope was created. The estimated population average time course of CD4+ cell depletion is 80.1857 CD4+ cells per milliliter of blood. The degree of heterogeneity across men in the rate of progression as time passes is 54.8061127978 cell count. The factors that predict cell count decline is time, pack of smoke, number of sexual partners, cesd mental illness score, age & time interaction, and smoke & time. The time factor is the most dramatic in term of CD4+ cell depletion.

1 Introduction

1.1 HIV and CD4+ Cells

Human immunodeficiency virus or HIV is a virus that attack immune system by killing a class of immune cell named CD4+ cell. On average a normal person without HIV have 1000 cells per milliliter of blood. As time passes from the initial HIV infection an infected person CD4+ cell counts starts to decline. Acquired immune deficiency syndrome or AIDS is the disease caused by the HIV virus.

1.2 The Data

The data used in this paper is a subset of the Multicenter AIDS Cohort Study with 369 men with HIV. The data consist of columns representing: time since seroconversion, CD4 count, age (relative to arbitrary origin), packs of cigarettes smoked per day, recreational drug use (yes/no), number of sexual partners, CESD (mental illness score), and subject ID. The data have been standardized, the measurements are unbalance, and the time interval are not evenly spaced.

1.3 Aim of the Investigation

The aim of the investigation is four main points: average time course of CD4+ cell depletion, time course for individual men, to characterize the degree of heterogeneity across men in the rate of progression, and factors which predict CD4+ cell changes.

2 Methods

2.1 Exploratory Data Analysis

The goal in exploratory data analysis (EDA) is to have an idea what the CD4+ cell count data looks like and ideas to go from EDA to modeling the data. Creating a response trend model will give an idea how time affect the response and if polynomial time is needed. A variogram graph will indicate what kind of variance is needed to be account for in the model. There are three different kind of variance either random effect variance, within-subject variance, and between-subject variance are needed.

2.2 Modeling Longitudinal Data

The next step is to create a suitable longitudinal model for the CD4+ cell data to answer the aim of this investigation. The model that will be chosen will have to address the variances that was shown in the variogram during EDA. After the model is selected the next step will be predictor selection. The predictor selection will be base on the deviance test of the full and the reduced model. Deviance test will be perform because the comparison are base on nested models.

2.3 Assumptions

The assumptions this investigation made is there are between-subject variations, within-subject variations, and measurement variations that need to be explicitly accounted for. The chosen longitudinal model will account for these explicitly so that the investigation can have an accurate and precise answers to the aim of this investigation.

Between-subject is latent factors. Latent factors are biological variability examples are diet, genetics, and other latent factors. Latent factors can keep an individuals CD4+ cell count consistently higher than the population mean or lower than the population mean.

The within-subject variation is serial correlation. The serial correlation is induced by time, the close two measurements are the more correlated they are. The farther apart two measurements are the less correlated they are.

Measurement variation takes into account for the process of taking measurements is an imperfect process and that there will be some variation in taking CD4+ cell count measurement. A variogram with force equally spacing of time intervals will confirm these assumptions of variations exist in the CD4+ cell count data.

3 Results

3.1 Exploratory Data Analysis Results

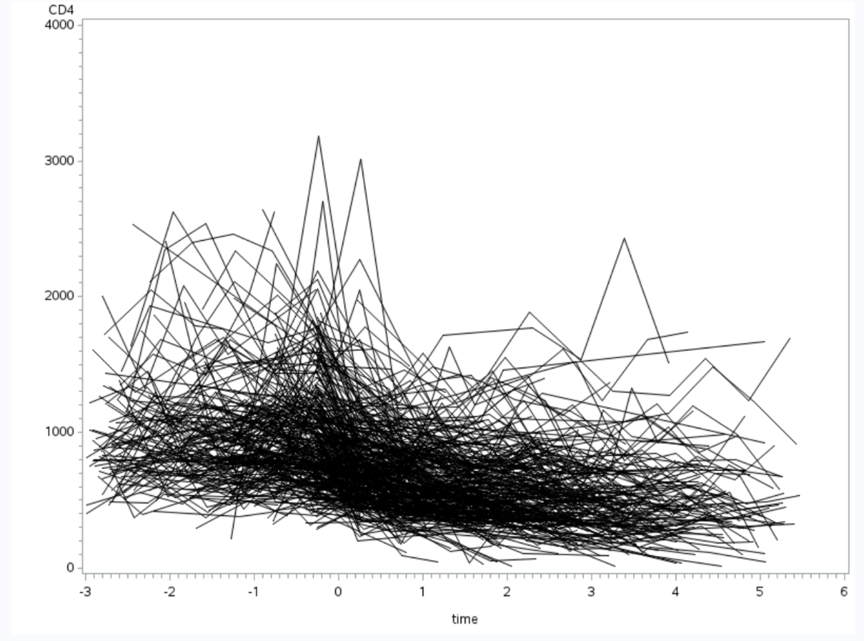


Figure 1: A graph between the response of the CD4+ cell count on the y-axis and the time points on the x-axis.

The spaghetti plot, Figure 1, shows that the data is unbalanced and that the time intervals are irregular and unequaled. It also show that individual have different base line which imply random intercept and that individual have different rate of progression which imply random slope. This will help in model selection especially when certain covariance structure have assumption about balance data and equally spaced time intervals.

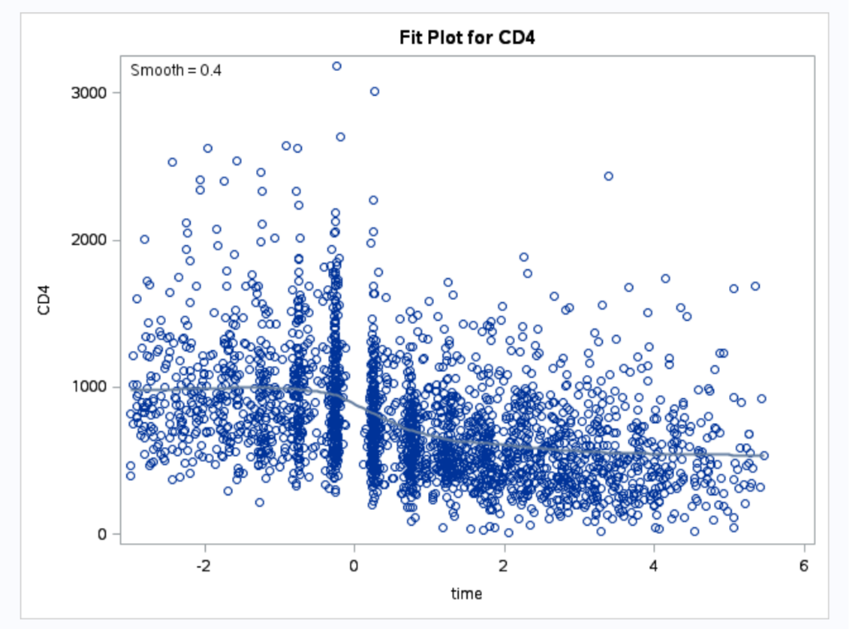


Figure 2: A graph between the response of the CD4+ cell count on the y-axis and the time points on the x-axis.

The response trend graph, Figure 2, indicate that perhaps time is not constant but some sort of polynomial. Between time point 0 and 2 months there is a sharp drop in CD4+ cell count and closer to the 2 month time point the CD4+ cell count rate of decline starts to steady out and the sharp decrease rate is slowed down drastically. Modeling the data with quadratic or cubic time predictor may be needed base on this graph.

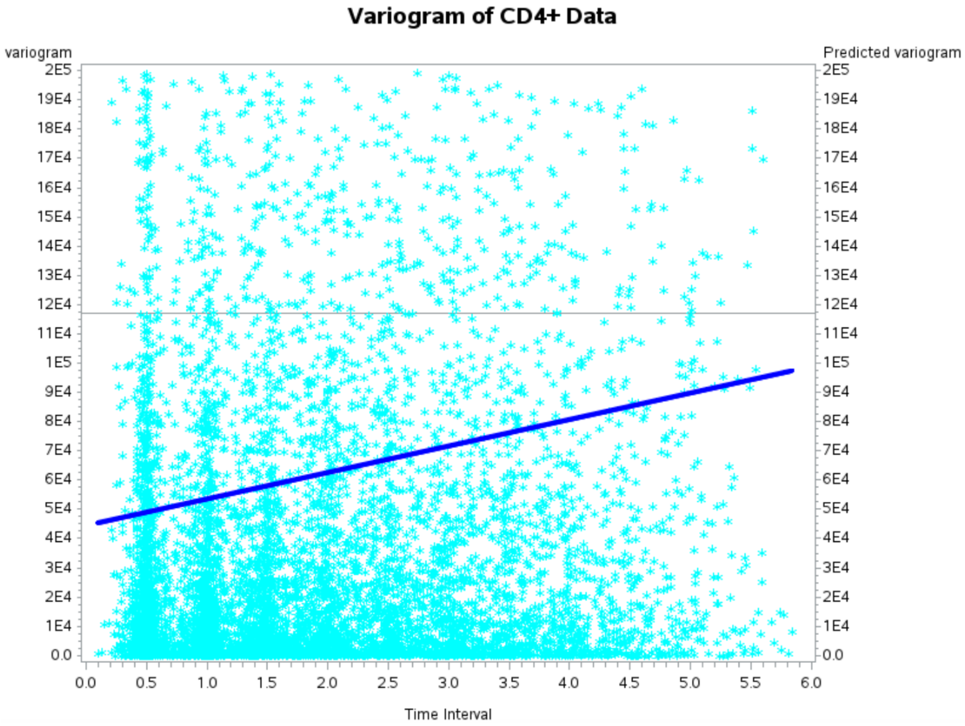


Figure 3: A variogram of the CD4+ cell count data with time intervals forced to be equally space.

Next is a plotted variogram (Figure 3) to check the assumption of having three sources of variation. Due to the data having unequaled time intervals the measurements are averaged and binned to the nearest time point. The blue line represent that variogram line and the grey horizontal line represents total variance.

Looking at Figure 3, the variogram blue solid line does not start at zero it indicate that there exist measurement errors. The variogram is not a flat blue line but a slanted line with a slope indicating that there exist serial correlation. Finally the blue line does not touched the upper limit of total variance indicating that there is random effect in play. The assumption that the CD4+ cell count data have all three sources of variation can be safely assume and is verified empirically.

3.2 Model Selection and Rejected Models

Longitudinal analysis have many linear models that to choose from. Models such as unstructured covariance and structured covariance. This section will discuss the reason for not choosing certain models.

Unstructured covariance is ruled out for two reasons. The first reason being that the large data set and large number of predictors would result in a large amount of parameter estimations. The second reason is that unstructured covariance is unsuitable for data set that have measurement taken at unequally spaced intervals.

Toeplitz covariance structure and autoregressive covariance structure both are other choices of structured covariance model. Both toeplitz and autoregressive assume that measurements are made at equal intervals of time. The CD4+ cell data have irregular unequal intervals of time.

The variogram shows there are three sources of variation. Independent model is rejected because the model assume there is only measurement error. Uniform model is also rejected because it only address two sources of variation, measurement error and between-individual variation. Exponential covariance model is rejected because the model address only within-individual variation.

Linear mixed effects models is chosen is because the model addresses all three sources of variation. The model explicitly distinguished between fixed and random effects. The advantage of this explicit distinction enable accurate and precise answers to the aim of this investigation.

3.3 Predictor Selection

Predictors	$\hat{\beta}$ values	p-values for t-test
intercept	790.11	<.0001
time	-81.6092	<.0001
age	1.6277	0.3790
smoke	41.0459	<.0001
drug	22.6537	0.2677
partners	6.5509	0.0043
cesd	-2.3499	0.0070
$age \times time$	-1.3805	0.0317
$smoke \times time$	-14.2323	<.0001
$drug \times time$	-1.7315	0.8488
$partners \times time$	-0.3958	0.7161
$cesd \times time$	0.1585	0.6899
$time^2$	0.8753	0.6187

Table 1: Full linear mixed effects model estimate.

After choosing the linear fixed effects model with random intercept and random slope to model the data, the next part is selecting a good combination of predictors that describe the CD4+ cell count data. A full model is fitted first. From Table 1, which show the estimated β, predictors that are not significant at p-value of 0.05 will be dropped and the predictors that are significant will be kept as a reduced model. Note the $time^2$ was included in the full model because of the nonlinear trend of time that was indicated in the response trend graph.

The predictors that are dropped are *drug*, $drug \times time$, $partners \times time$, $cesd \times time$, and $time^2$. Even though the *age* predictor is not significant the interaction $age \times time$ is significant therefore the *age* predictor is kept in the reduced model.

	Full Model	Reduced Model
-2 Log Likelihood	33603.4	33600.9
χ^2 Test Statistic	2.5	2.5
Degree of Freedom	13	8
$\chi^2_{5,0.95}$	11.070	11.070

Table 2: Likelihood Ratio test for two linear mixed effect models.

Hypothesis $\mathbf{H_1}$: Reduced Linear Mixed Effects Model

Hypothesis $\mathbf{H_2}$: Full Linear Mixed Effects Model

After fitting the reduced model, a likelihood ratio test was conducted between the full model and the reduced model. Table 2 shows the χ^2 test statistic at 2.5 which is the difference between the -2 Log Likelihood of full model and reduced model. The degree of freedom for χ^2 is the difference between the number of parameters in the full model and the number of parameters in the reduced model which is 5. The null hypothesis for the deviance test is the reduced model and the alternative hypothesis is the full model. Since the test statistic is 2.5 which is much less than 11.070, the reduced model is chosen.

3.4 Final Model

The equation listed below is the selected model that best represent the CD4+ cell count data and the best explanation of the data. With this model, the investigation can proceed to answer the aim of the investigation.

$$\begin{aligned}
Y_{ij} =& \beta_0 + \beta_1 \, time_{ij} + \beta_2 \, age_{ij} + \beta_3 \, smoke_{ij} + \beta_4 \, partners_{ij} + \beta_5 \, cesd_{ij} + \\
& \beta_6 \, age_{ij} \times time_{ij} + \beta_7 \, smoke_{ij} \times time_{ij} + b_{0i} + b_{1i} \times time_{ij} + e_{ij} \\
=& 791.05 - 80.1857 \, time_{ij} + 1.4697 \, age_{ij} + 38.0785 \, smoke_{ij} + 7.0434 \, partners_{ij} - \\
& 2.2867 \, cesd_{ij} - 1.3400 \, age_{ij} \times time_{ij} - 13.2674 \, smoke_{ij} \times time_{ij} + b_{0i} + \\
& b_{1i} \, time_{ij} + e_{ij}
\end{aligned} \tag{1}$$

Where b_{0i} represents the random intercept for each individual and b_{1i} represents the random slope for each individual.

The model can be rewritten in matrix notation

$$\underline{Y}_i = X_i \underline{\beta} + Z_i \underline{b}_i + \underline{e}_i, \quad i = 1, ..., N, \, j = 1, ..., n_i \tag{2}$$

where $\underline{Y}_i$ is a vector of size $n_i \times 1$ representing observations for ith individual, j represent the jth measurement for ith individual, X_i is a $n_i \times p$ design matrix of p independent fixed effect variables, Z_i is a $n_i \times q$ design matrix of q independent random effect variables, $\underline{\beta}$ is a vector of size $p \times 1$ representing fixed effect parameters, $\underline{b}_i$ is an independent vector of $q \times 1$ size representing random effects with $\mathcal{MVN}(\underline{0}, G)$ distribution (Multivariate Normal), and $\underline{e}_i$ represents an independent vector of random errors of size $n_i \times 1$ with $\mathcal{MVN}(\underline{0}, R_i)$ distribution. The $\underline{e}_i$ are independent of b_i.

The R_i represent within-subject variance. Linear fixed effects model break R_i down into two sources of within-subject variance, serial correlation and measurement error. The measurement error variance (τ^2) is equal to 59104. The serial correlation variance (σ^2) is 1.0649. The G matrix represents the between-subject variance.

$$\underline{\hat{\beta}} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \\ \hat{\beta}_3 \\ \hat{\beta}_4 \\ \hat{\beta}_5 \\ \hat{\beta}_6 \\ \hat{\beta}_7 \end{bmatrix} = \begin{bmatrix} 791.05 \\ -80.1857 \\ 1.4697 \\ 38.0785 \\ 7.0434 \\ -2.2867 \\ -1.3400 \\ -13.2674 \end{bmatrix}_{8\times 1}, \quad \underline{b}_i = \begin{bmatrix} b_{0i} \\ b_{1i} \end{bmatrix}_{2\times 1}, \quad \underline{b}_i \sim \mathcal{MVN}(\underline{0},\, G)$$

$$Var[\underline{b}_i] \;=\; G \;=\; \begin{bmatrix} Var[b_{0i}] & Cov[b_{0i}, b_{1i}] \\ Cov[b_{0i}, b_{1i}] & Var[b_{1i}] \end{bmatrix} = \begin{bmatrix} g_{11} & g_{12} \\ g_{21} & g_{22} \end{bmatrix} = \begin{bmatrix} 58244 & -3530.91 \\ -3530.91 & 3003.71 \end{bmatrix}_{2\times 2}$$

$$\underline{e}_i \;=\; \begin{bmatrix} e_{i1} \\ e_{i2} \\ \vdots \\ e_{in_i} \end{bmatrix}_{n_i\times 1}, \quad \underline{e}_i \sim \mathcal{MVN}(\underline{0},\, R_i)$$

$$R_i = Var[\underline{e}_i] = Var[\underline{e}_{i1}] + Var[\underline{e}_{i2}] = \sigma^2\Gamma + \tau^2\mathbb{I} = 1.0649\Gamma + 59104\,\mathbb{I}$$

3.5 Fixed Effect Interpretations and Significant

Predictors	$\hat{\beta}$ values	p-values for t-test
intercept	791.05	<.0001
time	-80.1857	<.0001
age	1.4697	0.4461
smoke	38.0785	<.0001
partners	7.0434	0.0009
cesd	-2.2867	0.0046
age × *time*	-1.3400	0.0369
smoke × *time*	-13.2674	<.0001

Table 3: Final model, reduced model, estimates.

Before interpretation of predictors start, as noted from previous section that the data are standardized. There are no statement on how the data is standardized. The age column, time column, and the number of sexual partners are clearly standardized. This will affect predictor interpretation in term of what one unit increase actually represents for each predictor therefore this investigation will intentionally leave out.

The general population of men with HIV starts out with an expected value of 791.05 CD4+ cells count per milliliter of blood. As time increases, from the baseline of HIV detection, the CD4+ cell per milliliter of blood decrease by 80.1857.

Age is not a significant predictor but the interaction with age and time is. Depending on an individual age and time progression affect the CD4+ count. Being at a certain age is not significant, it is the progression of time lag from the initial HIV detection with age that is significant. Mental illness base on cesd score decrease CD4+ cell count by 2.2867 per one unit increase, not as a steep decrease rate as time. The number of partners increases CD4+ cell count by 7.0434 per sexual partner.

What is surprising is that the number of smoking pack is a significant predictor for increasing CD4+ cell count. But this is negated by the interaction of time and smoking, the longer an individual smoke the more one decrease in CD4+ cell count. So smoking does not help when the interaction of time and smoking is taken into account for.

4 Conclusion

4.1 Aim of the Investigation

$$\begin{aligned} E[Y_{ij}] = {} & 791.05 \; - \; 80.1857\,time_{ij} \; + \; 1.4697\,age_{ij} \; + \; 38.0785\,smoke_{ij} \; + \; 7.0434\,partners_{ij} \; - \\ & 2.2867\,cesd_{ij} \; - \; 1.3400\,age_{ij} \times time_{ij} \; - \; 13.2674\,smoke_{ij} \times time_{ij} \end{aligned} \tag{3}$$

The estimated population average time course of CD4+ cell depletion is 80.1857 CD4+ cells per milliliter of blood. This highlight the natural progression of the disease AIDS caused by HIV as the time passes.

$$\begin{aligned} E[Y_{ij}|b_{0i}, b_{1i}] = {} & 791.05 \; - \; 80.1857\,time_{ij} \; + \; 1.4697\,age_{ij} \; + \; 38.0785\,smoke_{ij} \; + \\ & 7.0434\,partners_{ij} \; - \; 2.2867\,cesd_{ij} \; - \; 1.3400\,age_{ij} \times time_{ij} \; - \\ & 13.2674\,smoke_{ij} \times time_{ij} \; + \; b_{0i} \; + \; b_{1i}\,time_{ij} \end{aligned} \tag{4}$$

A few estimated time course for individual men.

ID	$\hat{\beta}_1$	$\hat{b}_{1i}$	$\hat{\beta}_1 + \hat{b}_{1i}$
10002	-80.1857	54.8061	-25.3796

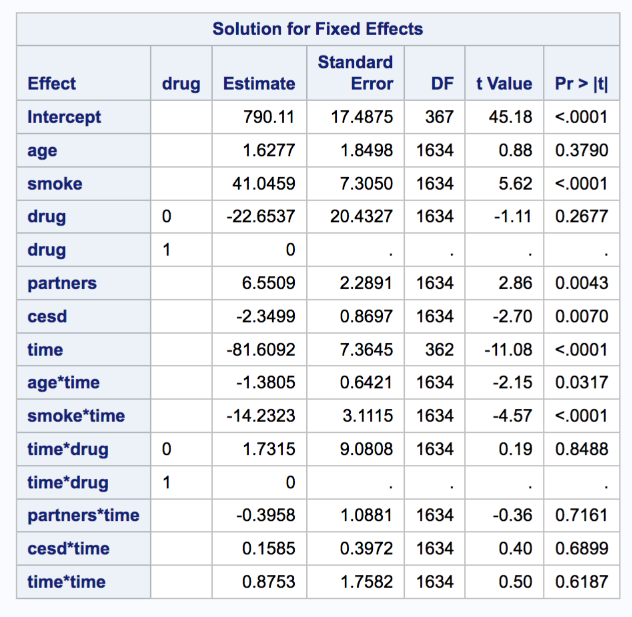
Table 4: A few estimated time course of CD4+ cells depletion for individual men.

The degree of heterogeneity across men in the rate of progression as time passes is characterized by the between-subject standard deviance which is square root of g_{22} value within the G matrix which is 54.8061127978 cell count. The rate of CD4+ cell decline difference between men is roughly 54.81 cells on top of the population rate of decline as time progress.

The factors that predict cell count decline is time, pack of smoke, number of sexual partners, cesd mental illness score, age & time interaction, and smoke & time. The time factor is the most dramatic in term of CD4+ cell depletion. As time progress the disease AIDS caused by HIV advances.

5 Appendix

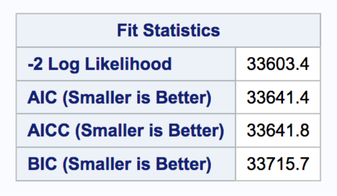
5.1 SAS Outputs



Solution for Fixed Effects

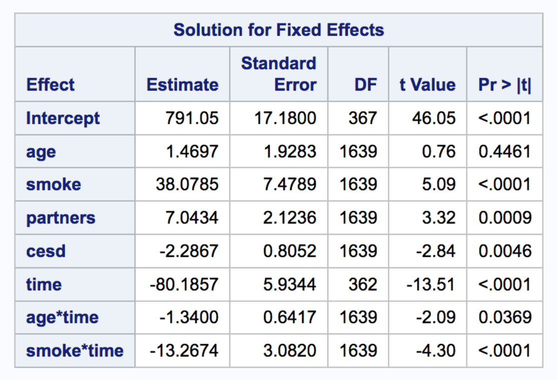
Effect	drug	Estimate	Standard Error	DF	t Value	Pr > |t|
Intercept		790.11	17.4875	367	45.18	<.0001
age		1.6277	1.8498	1634	0.88	0.3790
smoke		41.0459	7.3050	1634	5.62	<.0001
drug	0	-22.6537	20.4327	1634	-1.11	0.2677
drug	1	0	.	.	.	.
partners		6.5509	2.2891	1634	2.86	0.0043
cesd		-2.3499	0.8697	1634	-2.70	0.0070
time		-81.6092	7.3645	362	-11.08	<.0001
age*time		-1.3805	0.6421	1634	-2.15	0.0317
smoke*time		-14.2323	3.1115	1634	-4.57	<.0001
time*drug	0	1.7315	9.0808	1634	0.19	0.8488
time*drug	1	0	.	.	.	.
partners*time		-0.3958	1.0881	1634	-0.36	0.7161
cesd*time		0.1585	0.3972	1634	0.40	0.6899
time*time		0.8753	1.7582	1634	0.50	0.6187

Figure 4: Full linear mixed effects model with $\hat{\beta}$ estimates.



Fit Statistics	
-2 Log Likelihood	33603.4
AIC (Smaller is Better)	33641.4
AICC (Smaller is Better)	33641.8
BIC (Smaller is Better)	33715.7

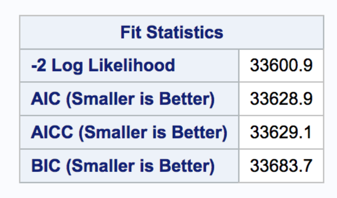
Figure 5: Full linear mixed effects model fit statistics.



Solution for Fixed Effects

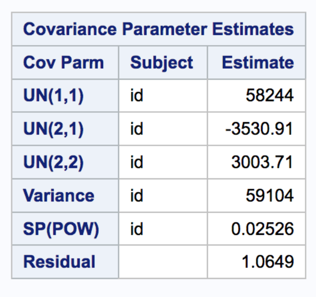
Effect	Estimate	Standard Error	DF	t Value	Pr > |t|
Intercept	791.05	17.1800	367	46.05	<.0001
age	1.4697	1.9283	1639	0.76	0.4461
smoke	38.0785	7.4789	1639	5.09	<.0001
partners	7.0434	2.1236	1639	3.32	0.0009
cesd	-2.2867	0.8052	1639	-2.84	0.0046
time	-80.1857	5.9344	362	-13.51	<.0001
age*time	-1.3400	0.6417	1639	-2.09	0.0369
smoke*time	-13.2674	3.0820	1639	-4.30	<.0001

Figure 6: Reduced linear mixed effects model with $\hat{\beta}$ estimates.



Fit Statistics	
-2 Log Likelihood	33600.9
AIC (Smaller is Better)	33628.9
AICC (Smaller is Better)	33629.1
BIC (Smaller is Better)	33683.7

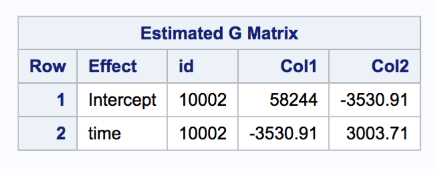
Figure 7: Reduced linear mixed effects model fit statistics.



Covariance Parameter Estimates

Cov Parm	Subject	Estimate
UN(1,1)	id	58244
UN(2,1)	id	-3530.91
UN(2,2)	id	3003.71
Variance	id	59104
SP(POW)	id	0.02526
Residual		1.0649

Figure 8: Reduced linear mixed effects model with estimated G and R_i matrices.



Estimated G Matrix

Row	Effect	id	Col1	Col2
1	Intercept	10002	58244	-3530.91
2	time	10002	-3530.91	3003.71

Figure 9: Estimated G matrix for subject with id 1002.

5.2 SAS Codes

```
proc import datafile='/home/mythicalprogramm0/my_courses/zhou_stat592/final_project/data/cd4.txt'
dbms=dlm out=cd4 replace;
delimiter='09'x;
getnames=yes;
run;

proc print data=cd4;
run;

symbol i=join repeat=369 color=black;
proc gplot data=Cd4;
plot CD4*time=id;
run;


/*Performing Loess smoothing*/
proc loess data=CD4;
model CD4=time/smooth=0.1 0.25 0.4 0.6;
symbol1 color=black value=dot;
symbol2 color=black interpol=join value=none;
run;


%include '/home/mythicalprogramm0/my_courses/zhou_stat592/final_project/macros/autocor.sas';

%autocor(data=cd4, y=CD4,time=time,id=id);

%include '/home/mythicalprogramm0/my_courses/zhou_stat592/final_project/macros/variogram.sas';


%variogram (data=cd4,resvar=cd4,clsvar=, expvars=time age smoke drug partners cesd time*age
time*smoke time*drug time*partners time*cesd time*time time*time*time,id=id,time=time,maxtime=12);


%include '/home/mythicalprogramm0/my_courses/zhou_stat592/final_project/macros/variance.sas';


%variance(data=cd4,id=id,resvar=cd4,clsvar=, expvars=time age smoke drug partners cesd time*age
time*smoke time*drug time*partners time*cesd time*time time*time*time,subjects=369,maxtime=12);


proc loess data=varioplot;
model variogram=time_interval;
ods output outputstatistics=stat;
run;

goptions reset=all;
proc gplot data=stat;
plot depvar*time_interval / vaxis=axis1 haxis=axis2 vref=117100;
plot2 pred*time_interval / vaxis=axis1 haxis=axis2;
symbol value=star color=cyan;
/*symbol2 v=none i=sm90s color=blue width=3;*/
symbol2 v=none  color=blue width=3 interpol=sm5s;
axis1 order=0 to 200000 by 10000;
axis2 order=0 to 6 by .5;
label time_interval='Time Interval';
format time_interval f3.1 depvar f4.1 pred f4.1;
title 'Variogram of CD4+ Data';
run;
quit;
```

```
data stat;
set stat;
autocorr=1-(pred/117100);
run;
goptions reset=all;
proc gplot data=stat;
plot autocorr*time_interval / vaxis=axis1 haxis=axis2;
symbol v=none i=sm60s;
axis1 order=0 to 1 by 0.1;
axis2 order=0 to 6 by .5;
label time_interval='Time Interval';
format time_interval f3.1 autocorr f4.1;
title 'Autocorrelation Plot of CD4+ Data';
run;
quit;
proc import datafile='/home/mythicalprogramm0/my_courses
/zhou_stat592/final_project/data/cd4.txt' dbms=dlm out=cd4 replace;
delimiter='09'x;
getnames=yes;
run;

proc mixed data=cd4 method=ml;
TITLE 'Full Model for CD4+ random intercept & random slope';
class drug id;
model cd4 = age smoke drug partners cesd time age*time smoke*time drug*time
partners*time cesd*time time*time /solution;
random intercept time/type=un subject=id g gcorr v vcorr;
repeated/type=sp(pow)(time) local subject=id r rcorr;
run;

PROC MIXED data=cd4 method=ml;
  TITLE 'Reduced Model for CD4+ random intercept & random slope';
  CLASS id drug;
  model cd4 = age smoke partners cesd time age*time smoke*time/solution;
  RANDOM INTERCEPT time/TYPE=un SUBJECT=id g gcorr v vcorr;
  REPEATED /TYPE=SP(POW)(time) LOCAL SUBJECT=id r rcorr;
RUN;
```