STAT 6225, Final Assignment
Computation Part (a)

```
> library(npmlda)
> x = read.csv("/Users/Aaron_Zhang/Desktop/NGHS.csv")
> summary(x)
> x= x[!is.na(x$BMI),]
> x = data.frame(x, agebin=numeric(nrow(x)))
> for (i in 9:18){
+   print(i)
+   for (j in 1:10){
+     x$agebin[( x$AGE*10 >= (i*10+ (j-1)) &  x$AGE*10 < (i*10+ j))]  =   (i*10+ (j-1))
+     print(c(i+ (j-1)/10))
+   } }
> x$agebin= x$agebin/10
> x$agebin= x$agebin/10
> x.b = x[x$RACE==2,]
> x.w = x[x$RACE==1,]
> ID.b = unique(x.b$ID)
> ID.w = unique(x.w$ID)
> nID.b = length(ID.b)
> nID.w = length(ID.w)
> grid = .2
> s1cat = seq(9.0, 18.8, by=grid)
> s12cat = seq(9.0, 16.8, by=grid)
> kk1 = length(s1cat)
> kk2 = length(s12cat)
> x.prob = function(data){
+   yvec = 1*(data$BMIPCT>=85)
+   xvec = data$agebin
+   IDD = data$ID
+   prob.s1 = numeric(kk1)
+   for (k in 1:kk1){
+     prob.s1[k] = NW.WtKernel(xvec,yvec,s1cat[k], Bndwdth=2.6)
+   }
+   prob.s12 = numeric(kk2)
+   for (k in 1:kk2){
+     prob.s12[k] = Kernel2D(IDD,xvec,yvec, X01=s12cat[k],X02=s12cat[k]
+2,Bndwdth1=0.9,Bndwdth2=0.9)
+   }
+   IND1 = (s1cat >=9 & s1cat <=16.8)
+   IND2 = (s1cat >=11 & s1cat <=18.8)
+   RTP = prob.s12/(prob.s1[IND1])
+   RTPR = RTP/(prob.s1[IND2])
+   results = as.list(NA)
+   results[[1]] = prob.s1
+   results[[2]] = prob.s12
+   results[[3]] = RTP
+   results[[4]] = RTPR
+   return(results)
+ }
> b.0 = x.prob(x.b)
> w.0 = x.prob(x.w)
> myboot.b = function(B){
+   results = as.list(rep(NA,B))
+   for (i in 1:B){
+     sample.ID = sample(ID.b, length(ID.b), replace = TRUE)
```

```
+     data = NA
+     for (j in 1:length(sample.ID)){
+       sample = x.b[x.b$ID==sample.ID[j],]
+       data = rbind(data,sample)
+     }
+     newdata = data[-1,]
+     results[[i]] = x.prob(newdata)
+   }
+   return(results)
+ }
> myboot.w = function(B){
+   results = as.list(rep(NA,B))
+   for (i in 1:B){
+     sample.ID = sample(ID.w, length(ID.w), replace = TRUE)
+     data = NA
+     for (j in 1:length(sample.ID)){
+       sample = x.w[x.w$ID==sample.ID[j],]
+       data = rbind(data,sample)
+     }
+     newdata = data[-1,]
+     results[[i]] = x.prob(newdata)
+   }
+   return(results)
+ }
> set.seed(123)
> results.b = myboot.b(500)
> results.w = myboot.w(500)
```

1.1

```
> b.s1.low = rep(NA,50)
> b.s1.up = rep(NA,50)
> results.b.1 = matrix(ncol = 50, nrow = 500)
> for (i in 1:50){
+   for (j in 1:500){
+     results.b.1[j,] = results.b[[j]][[1]]
+   }
+   b.s1.low[i] = quantile(results.b.1[,i],0.025)
+   b.s1.up[i] = quantile(results.b.1[,i],0.975)
+ }
> w.s1.low = rep(NA,50)
> w.s1.up = rep(NA,50)
> results.w.1 = matrix(ncol = 50, nrow = 500)
> for (i in 1:50){
+   for (j in 1:500){
+     results.w.1[j,] = results.w[[j]][[1]]
+   }
+   w.s1.low[i] = quantile(results.w.1[,i],0.025)
+   w.s1.up[i] = quantile(results.w.1[,i],0.975)
+ }
```

1.2

```
> b.s12.low = rep(NA,40)
> b.s12.up = rep(NA,40)
> results.b.2 = matrix(ncol = 40, nrow = 500)
```

```
> for (i in 1:40){
+   for (j in 1:500){
+     results.b.2[j,] = results.b[[j]][[2]]
+   }
+   b.s12.low[i] = quantile(results.b.2[,i],0.025)
+   b.s12.up[i] = quantile(results.b.2[,i],0.975)
+ }
> w.s12.low = rep(NA,40)
> w.s12.up = rep(NA,40)
> results.w.2 = matrix(ncol = 40, nrow = 500)
> for (i in 1:40){
+   for (j in 1:500){
+     results.w.2[j,] = results.w[[j]][[2]]
+   }
+   w.s12.low[i] = quantile(results.w.2[,i],0.025)
+   w.s12.up[i] = quantile(results.w.2[,i],0.975)
+ }
```

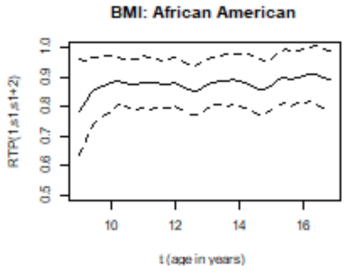
2.1

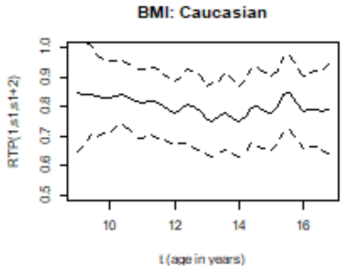
```
> b.rtp.low = rep(NA,40)
> b.rtp.up = rep(NA,40)
> results.b.3 = matrix(ncol = 40, nrow = 500)
> for (i in 1:40){
+   for (j in 1:500){
+     results.b.3[j,] = results.b[[j]][[3]]
+   }
+   b.rtp.low[i] = quantile(results.b.3[,i],0.025)
+   b.rtp.up[i] = quantile(results.b.3[,i],0.975)
+ }
> w.rtp.low = rep(NA,40)
> w.rtp.up = rep(NA,40)
> results.w.3 = matrix(ncol = 40, nrow = 500)
> for (i in 1:40){
+   for (j in 1:500){
+     results.w.3[j,] = results.w[[j]][[3]]
+   }
+   w.rtp.low[i] = quantile(results.w.3[,i],0.025)
+   w.rtp.up[i] = quantile(results.w.3[,i],0.975)
+ }
```

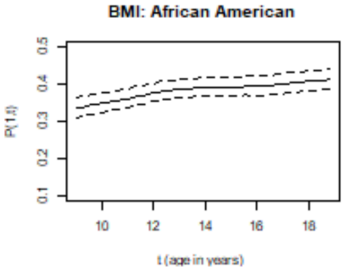
2.2

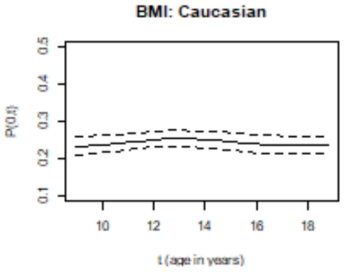
```
> b.rtpr.low = rep(NA,40)
> b.rtpr.up = rep(NA,40)
> results.b.4 = matrix(ncol = 40, nrow = 500)
> for (i in 1:40){
+   for (j in 1:500){
+     results.b.4[j,] = results.b[[j]][[4]]
+   }
+   b.rtpr.low[i] = quantile(results.b.4[,i],0.025)
+   b.rtpr.up[i] = quantile(results.b.4[,i],0.975)
+ }
> w.rtpr.low = rep(NA,40)
> w.rtpr.up = rep(NA,40)
> results.w.4 = matrix(ncol = 40, nrow = 500)
```

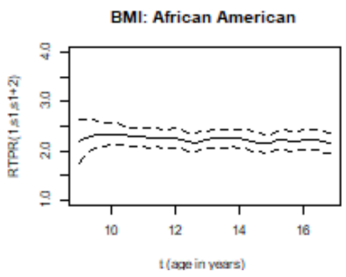
```
> for (i in 1:40){
+   for (j in 1:500){
+     results.w.4[j,] = results.w[[j]][[4]]
+   }
+   w.rtpr.low[i] = quantile(results.w.4[,i],0.025)
+   w.rtpr.up[i] = quantile(results.w.4[,i],0.975)
+ }
> par(mfrow=c(2,2),cex=0.5, mai=c(.4,.4,.4,.4))
> plot(s1cat, b.0[[1]], type = "l", ylim = c(0.1,0.5),
+      xlab = "t (age in years)", ylab = "P(1,t)", main = "BMI: African American")
> lines(s1cat,b.s1.low, lty=2)
> lines(s1cat,b.s1.up, lty=2)
> plot(s1cat, w.0[[1]], type = "l", ylim = c(0.1,0.5),
+      xlab = "t (age in years)", ylab = "P(0,t)", main = "BMI: Caucasian")
> lines(s1cat,w.s1.low, lty=2)
> lines(s1cat,w.s1.up, lty=2)
> plot(s12cat, b.0[[2]], type = "l", ylim = c(0.1,0.5),
+      xlab = "t (age in years)", ylab = "P(1,s1,s1+2)", main = "BMI: African American")
> lines(s12cat,b.s12.low, lty=2)
> lines(s12cat,b.s12.up, lty=2)
> plot(s12cat, w.0[[2]], type = "l", ylim = c(0.1,0.5),
+      xlab = "t (age in years)", ylab = "P(0,s1,s1+2)", main = "BMI: Caucasian")
> lines(s12cat,w.s12.low, lty=2)
> lines(s12cat,w.s12.up, lty=2)
> par(mfrow=c(2,2),cex=0.5, mai=c(.4,.4,.4,.4))
> plot(s12cat, b.0[[3]], type = "l", ylim = c(0.5,1),
+      xlab = "t (age in years)", ylab = "RTP(1,s1,s1+2)", main = "BMI: African American")
> lines(s12cat,b.rtp.low, lty=2)
> lines(s12cat,b.rtp.up, lty=2)
> plot(s12cat, w.0[[3]], type = "l", ylim = c(0.5,1),
+      xlab = "t (age in years)", ylab = "RTP(1,s1,s1+2)", main = "BMI: Caucasian")
> lines(s12cat,w.rtp.low, lty=2)
> lines(s12cat,w.rtp.up, lty=2)
> plot(s12cat, b.0[[4]], type = "l", ylim = c(1,4),
+      xlab = "t (age in years)", ylab = "RTPR(1,s1,s1+2)", main = "BMI: African American")
> lines(s12cat,b.rtpr.low, lty=2)
> lines(s12cat,b.rtpr.up, lty=2)
> plot(s12cat, w.0[[4]], type = "l", ylim = c(1,4),
+      xlab = "t (age in years)", ylab = "RTPR(0,s1,s1+2)", main = "BMI: Caucasian")
> lines(s12cat,w.rtpr.low, lty=2)
> lines(s12cat,w.rtpr.up, lty=2)
```

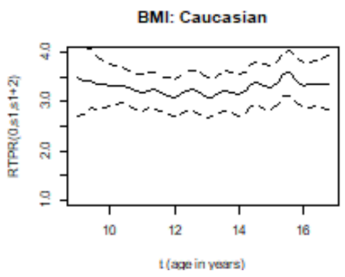


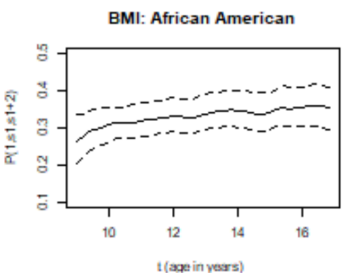


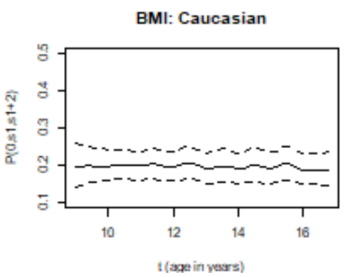












STAT 6225, Final Assignment

Writing Part (b)

Abstract

The research on AIDS has always been a hot topic in the medical field. In order to let AIDS patients live a better life, in this paper, we will use the data from BMACS(Baltimore site of the Multicenter AIDS Cohort Study) to explore whether several factors affect the percentage of CD4 lymphocyte consumption after infection. After using five models for analysis, we find that preCD4 is significant but smoke is insignificant and we found the best model for each of these 5 models (except time-varying coefficient model). After that, we will summarize our findings and discuss the strengths, weaknesses, and applicability of each model.

Keywords: AIDS; BMACS; CD4; Smoke

1 Introduction

Human immunodeficiency virus infection and acquired immunodeficiency syndrome (HIV/AIDS) is a spectrum of conditions caused by infection with the human immunodeficiency virus (HIV). Following initial infection a person may not notice any symptoms, or may experience a brief period of influenza-like illness. Typically, this is followed by a prolonged period with no symptoms. If the infection progresses, it interferes more with the immune system, increasing the risk of developing common infections such as tuberculosis, as well as other opportunistic infections, and tumors which are otherwise rare in people who have normal immune function. These late symptoms of infection are referred to as acquired immunodeficiency syndrome (AIDS). This stage is often also associated with unintended weight loss.

HIV is spread primarily by unprotected sex (including anal and oral sex), contaminated blood transfusions, hypodermic needles, and from mother to child during pregnancy, delivery, or breastfeeding. Some bodily fluids, such as saliva, sweat and tears, do not transmit the virus. HIV is a member of the group of viruses known as retroviruses.

Due to its transmission characteristics, AIDS has become one of the most serious diseases in the world. In the medical community, the research on AIDS has never stopped.

In this paper, we will explore how the time after infection, cigarette smoking status, patients' age, and relative CD4 cell levels before the infection will have influence on the post-infection depletion of CD4 percent of lymphocytes.

2 Data and objectives

This dataset is from the Baltimore site of the Multi-center AIDS Cohort Study (BMACS), which included 400 homosexual men who were infected by the human immunodeficiency virus (HIV) between 1984 and 1991.

Details:

ID. Subject ID

Time. Subject's study visit time

Smoke. Cigarette baseline smoking status

Age. Age at study enrollment

preCD4. Pre-infection CD4 percentage

CD4. CD4 percentage at the time of visit

Although all the individuals were scheduled to have their measurements made at semi-annual visits, the study has an unbalanced design because the subjects' actual

visiting times did not exactly follow the schedule and the HIV infections happened randomly during the study.

The covariates of interest in these data include both time-dependent and time-invariant variables. Details of the statistical design and scientific importance of the BMACS data can be found in Kaslow et al. (1987) and Wu, Chiang and Hoover (1998). The BMACS data used in this book included 283 subjects with a total of 1817 observations.

We will explore how these variables affect the percentage of CD4 cells and establish models to explore the relationships between these variables in the following chapters.

3 Preliminary Analysis

First, we can find that the data set is longitude data, and the same ID represents the same patient, but the same person is not independent of each other and the standard linear model needs independence between those error terms while mixed-effects method provides a new approach for the analysis of the data set, i.e., it assumes that the conditional independence so as to help in these data sets processing related issues, such as BMACS data and then we can apply mixed-effects model. For the variables in the dataset, we will only consider the variables Smoke, age, preCD4, and CD4 to see if they are evenly distributed in this part.

For variable Smoke, we can see that it has two factors, respectively is 0 and 1 and there should not be any relationship between each patient, and then image shows there are more than 600 smokers and 1200 non-smokers, this could mean that nonsmokers are more likely to be infected with HIV, or the study was conducted in the case of unbalanced.

For the variable age, it ranges from 18 to 60 years old and the graph shows that it is normally distributed in some way, which means that this variable doesn't need any transformation.

For the variable preCD4, it ranges from 15 to 69 and the graph shows that it is normally distributed, which means that this variable doesn't need any transformation.

For the variable CD4, it ranges from 1 to 63 and the graph shows that it is normally distributed, which means that this variable doesn't need any transformation.

We also need to check the data before we start the analysis. Because of the large number of patients, we selected 18 smokers and 18 non-smokers respectively as samples, and then stored each sample as a grouping data object. Packet data objects are enhanced data frames provided by the NLME package, which incorporate model formulas that give information about the data structure. In the formula CD4

~Time | ID, CD4 is the response variable, Time is the principal within-group covariate, and ID is the grouping variable.

By analyzing the scatter plots drawn by the above two models, we found that the percentage of CD4 cells after infection by most non-smokers was weakly negatively correlated with time, but there was a difference among different patients, that is, the slope of regression curve of some patients was close to zero or even positive. For most smokers, there is also a negative correlation between the two variables, where the average slope is higher. Considering that each patient had fewer time points, linear regression was able to reasonably fit the relationship between the percentage of CD4 cells and time after infection, regardless of whether the patient smoked.

4 Statistical Models and Methods

4.1 Linear mixed-effects model

4.1.1 Simple linear mixed-effects model

In question 1 of assignment 1, we get

$$Y_{ij} = \beta_{0i} + \beta_{1i}T_{ij} + \varepsilon_{ij}$$

Then we can get 9 models:

m1: $\beta_{0i} = \gamma_{00} + \gamma_{01}S_i + u_{0i}$, $\beta_{1i} = \gamma_{10} + \gamma_{11}S_i$

m2: $\beta_{0i} = \gamma_{00} + \gamma_{01}P_i + u_{0i}$, $\beta_{1i} = \gamma_{10} + \gamma_{11}P_i$

m3: $\beta_{0i} = \gamma_{00} + \gamma_{01}S_i + \gamma_{02}P_i + u_{0i}$, $\beta_{1i} = \gamma_{10} + \gamma_{11}S_i + \gamma_{12}P_i$

m4: $\beta_{0i} = \gamma_{00} + \gamma_{01}S_i$, $\beta_{1i} = \gamma_{10} + \gamma_{11}S_i + u_{1i}$

m5: $\beta_{0i} = \gamma_{00} + \gamma_{01}P_i$, $\beta_{1i} = \gamma_{10} + \gamma_{11}P_i + u_{1i}$

m6: $\beta_{0i} = \gamma_{00} + \gamma_{01}S_i + \gamma_{02}P_i$, $\beta_{1i} = \gamma_{10} + \gamma_{11}S_i + \gamma_{12}P_i + u_{1i}$

m7: $\beta_{0i} = \gamma_{00} + \gamma_{01}S_i + u_{0i}$, $\beta_{1i} = \gamma_{10} + \gamma_{11}S_i + u_{1i}$

m8: $\beta_{0i} = \gamma_{00} + \gamma_{01}P_i + u_{0i}$, $\beta_{1i} = \gamma_{10} + \gamma_{11}P_i + u_{1i}$

m9: $\beta_{0i} = \gamma_{00} + \gamma_{01}S_i + \gamma_{02}P_i + u_{0i}$, $\beta_{1i} = \gamma_{10} + \gamma_{11}S_i + \gamma_{12}P_i + u_{1i}$

4.1.2 Polynomial linear mixed-effects model

In question 2 of assignment 1, we get

$$Y_{ij} = \beta_{0i} + \beta_{1i}T_{ij} + \beta_{2i}T_{ij}^{\,2} + \varepsilon_{ij}$$

Then we can get 12 models:

lm1: $\beta_{0i} = \gamma_{00} + \gamma_{01}S_i + u_{0i}$, $\beta_{1i} = \gamma_{10} + \gamma_{11}S_i$, $\beta_{2i} = \gamma_{20} + \gamma_{21}S_i$

lm2: $\beta_{0i} = \gamma_{00} + \gamma_{01}P_i + u_{0i}$, $\beta_{1i} = \gamma_{10} + \gamma_{11}P_i$, $\beta_{2i} = \gamma_{20} + \gamma_{21}P_i$

lm3: $\beta_{0i} = \gamma_{00} + \gamma_{01}S_i + \gamma_{02}P_i + u_{0i}$, $\beta_{1i} = \gamma_{10} + \gamma_{11}S_i + \gamma_{12}P_i$, $\beta_{2i} = \gamma_{20} + \gamma_{21}S_i + \gamma_{22}P_i$

lm4: $\beta_{0i} = \gamma_{00} + \gamma_{01}S_i$, $\beta_{1i} = \gamma_{10} + \gamma_{11}S_i + u_{1i}$, $\beta_{2i} = \gamma_{20} + \gamma_{21}S_i$

lm5: $\beta_{0i} = \gamma_{00} + \gamma_{01}P_i$, $\beta_{1i} = \gamma_{10} + \gamma_{11}P_i + u_{1i}$, $\beta_{2i} = \gamma_{20} + \gamma_{21}P_i$

lm6: $\beta_{0i} = \gamma_{00} + \gamma_{01}S_i + \gamma_{02}P_i$, $\beta_{1i} = \gamma_{10} + \gamma_{11}S_i + \gamma_{12}P_i + u_{1i}$, $\beta_{2i} = \gamma_{20} + \gamma_{21}S_i + \gamma_{22}P_i$

lm7: $\beta_{0i} = \gamma_{00} + \gamma_{01}S_i + u_{0i}$, $\beta_{1i} = \gamma_{10} + \gamma_{11}S_i + u_{1i}$, $\beta_{2i} = \gamma_{20} + \gamma_{21}S_i$

lm8: $\beta_{0i} = \gamma_{00} + \gamma_{01}P_i + u_{0i}$, $\beta_{1i} = \gamma_{10} + \gamma_{11}P_i + u_{1i}$, $\beta_{2i} = \gamma_{20} + \gamma_{21}P_i$

lm9: $\beta_{0i} = \gamma_{00} + \gamma_{01}S_i + \gamma_{02}P_i + u_{0i}$, $\beta_{1i} = \gamma_{10} + \gamma_{11}P_i + \gamma_{12}P_i + u_{1i}$, $\beta_{2i} = \gamma_{20} + \gamma_{21}S_i$

$+ \gamma_{22}P_i$

lm10: $\beta_{0i} = \gamma_{00} + \gamma_{01}S_i + u_{0i}$, $\beta_{1i} = \gamma_{10} + \gamma_{11}S_i + u_{1i}$, $\beta_{2i} = \gamma_{20} + \gamma_{21}S_i + u_{2i}$

lm11: $\beta_{0i} = \gamma_{00} + \gamma_{01}P_i + u_{0i}$, $\beta_{1i} = \gamma_{10} + \gamma_{11}P_i + u_{1i}$, $\beta_{2i} = \gamma_{20} + \gamma_{21}P_i + u_{2i}$

lm12: $\beta_{0i} = \gamma_{00} + \gamma_{01}S_i + \gamma_{02}P_i + u_{0i}$, $\beta_{1i} = \gamma_{10} + \gamma_{11}P_i + \gamma_{12}P_i + u_{1i}$, $\beta_{2i} = \gamma_{20} + \gamma_{21}S_i$

$+ \gamma_{22}P_i + u_{2i}$

4.2 Nonparametric model

Kernel estimator of $\mu(t) = E(Y(t) \mid t)$ can minimize the local score function

$$\ell_K\left(t; h, N^{-1}\right) = \frac{1}{N}\sum_{i=1}^{n}\sum_{j=1}^{n_i}\left[Y_{ij} - \mu(t)\right]^2 K\left(\frac{t - t_{ij}}{h}\right)$$

Then we get

$$\ell_K(t; h, w) = \sum_{i=1}^{n}\sum_{j=1}^{n_i} w_i\left[Y_{ij} - \mu(t)\right]^2 K\left(\frac{t - t_{ij}}{h}\right)$$

Then, the kernel estimator:

$$\widehat{\mu}_K\left(t; h, w^*\right) = \frac{\sum_{i=1}^{n}\left\{n_i^{-1}\sum_{j=1}^{n_i} Y_{ij} K\left[(t - t_{ij})/h\right]\right\}}{\sum_{i=1}^{n}\left\{n_i^{-1}\sum_{j=1}^{n_i} K\left[(t - t_{ij})/h\right]\right\}}$$

The next step is resampling-subject bootstrap:

(a) Denote by the longitudinal sample: $\left\{\left(Y_{ij}^*, t_{ij}^*\right) : i = 1, \ldots, n; j = 1, \ldots, n_i\right\}$

(b) Compute the curve estimator: $\widehat{\mu_K^{\text{boot}}}(t; h, w)\ of\ \mu(t)$

(c) Repeat the above two steps B times and get B bootstrap estimator:

$$\mathscr{B}_\mu^B(t; h, w) = \left\{\widehat{\mu}_L^{boot,1}(t; h, w), \ldots, \widehat{\mu}_L^{boot,B}(t; h, w)\right\}$$

4.3 Piece-wise linear mixed-effects model

B-Splines Least Squares Criteria:

Suppose that each $\beta_l(t)$, $l = 0, \ldots, k$, can be approximated by a set of B- spline basis functions $\mathscr{B}_l^{(S)}(t) = \left\{B_{ls}^{(S)}(t) : t \in \mathscr{T}; s = 1, \ldots, K_l\right\}$, which spans the linear space

$$\mathbb{G}_l^{(S)} = \left\{\text{ linear function space spanned by } \mathscr{B}_l^{(S)}(t)\right\}.$$

Once $\mathscr{B}_l^{(S)}(t)$ is selected with a fixed K_l, $\mathbb{G}_l^{(S)}$ is a linear space with a fixed degree and knot sequence. We then have the B-spline basis approximation

$$\beta_l(t) \approx \sum_{k=1}^{K_l}\left[\gamma_{lk}^* B_{ls}^{(S)}(t)\right].$$

B-spline approximated time varying coefficient model:

$$\begin{cases} Y_{ij} \approx Y_{ij}^{(S)}, \\ Y_{ij}^{(S)} = \sum_{l=0}^{k} \sum_{s=1}^{K_l} \left[X_{ij}^{(l)} \gamma_{ls}^{*} B_{ls}^{(S)}(t_{ij}) \right] + \varepsilon_{ij}, \end{cases}$$

B-spline linear model: $\sum_{l=0}^{k} \sum_{s=1}^{K_l} \left[X_{ij}^{(l)} \gamma_{ls}^{*} B_{ls}^{(S)}(t_{ij}) \right]$.

B-spline coefficients:

$$\ell_S(\gamma) = \sum_{i=1}^{n} \sum_{j=1}^{n_i} \left\{ w_i \left[Y_{ij} - \sum_{l=0}^{k} \sum_{s=1}^{K_l} X_{ij}^{(l)} B_{lk}^{(S)}(t_{ij}) \gamma_{lk} \right]^2 \right\}$$

B-spline coefficients (matrix expression):

$$\ell_S(\gamma) = \sum_{i=1}^{n} \left[(Y_i - \mathbf{U}_i^{(S)} \gamma)^T \mathbf{W}_i (Y_i - \mathbf{U}_i^{(S)} \gamma) \right].$$

5 Findings and Interpretations

5.1 Assignment 1

5.1.1 Simple linear mixed-effect model

Through the previous preliminary analysis, we show the necessity of using the mixed effect model to avoid the non-independence between the same individuals. As we discussed in the previous section, there are nine simple linear mixed-effect models. We then want to compare the different types of models separately, but the anova test can only be applied to two models with the same fixed effect terms, which means that for models with different fixed effect terms, we need to use the AIC and compare the values to determine which model is better. Since the best model of each model has the same fixed effect term, we found that the second

model of the 9 models by anova test is the best model of the simple linear mixed effect model, i.e., if there were no polynomial function in the model, we might think that smoking status had little effect on the percentage of CD4 cells,

5.1.2 Polynomial linear mixed-effects model

Then, in the next 12 models, we also found that each of the best models had the same fixed effect term. Therefore, through the test of variance analysis, we find that the 11th model of the polynomial linear mixed-effects models has the best effect, i.e., in the polynomial function of this model, we may think that smoking status still does not have a great influence on the results, but for the second polynomial term, it is necessary to consider random intercepts and random coefficients with random slopes.

Finally, we use AIC again and find that the optimal polynomial linear mixed effect model is superior to the simple linear mixed effect model. Also, we found that the variable time had a more significant impact on CD4 variable, and preCD4 also had a great impact on the response variable.

5.2 Assignment 2

5.2.1 Simple nonparametric model

We will use simple nonparametric model of post-infection CD4 as a function of time without incorporating the covariate effects and we will find the mean and confidence interval of the curve of CD4 percentage over time.

We find that the spline estimator is better than other two uniform estimators for simple nonparametric models and the curve is smoother when there is a larger bandwidth for the uniform estimator, i.e., the cubic spline estimator is best among the three estimators.

5.2.2 Time-varying coefficient model

We compared uniform estimate coefficients with bandwidth between 0.5 and 1 for time-varying coefficient models and found that the larger the bandwidth, the smoother the coefficients. Also, we found that the coefficients in the spline model is smoother than the other two but it is not obvious. Therefore, we can conclude that there are little difference between those models, and it is difficult to compare the changes of these coefficients in 5 years. In general, the effect of smoke will increase but the effect of preCD4 and age will decrease as year increases, i.e., smoking status is important for patients with long-term illnesses, which may need to be controlled by doctors, while the other two are important for patients with short-term illnesses.

5.3 Assignment 3

5.3.1 Piece-wise linear mixed-effects model

The local constant fitting using the basic function model is compared with the local linear fitting without assuming the continuity of the node model and we know that local linear fitting is better without assuming continuity of node model. We analyze the piece-wise linear mixed-effects models with different basis functions by using anova function and we find that the third model is the best, i.e., the local linear fit model with continuity at the knots is the best. Then, we use bootstrap to find the confidence interval for each estimator in this model and we can see that though the influence fluctuates but it still changes in a relative small level.

5.3.2 Varying-coefficient mixed-effect model

The local constant fitting using the basic function model is compared with the local linear fitting without assuming the continuity of the node model and we know that local linear fitting is better without assuming continuity of node model. We analyze the varying-coefficient mixed-effects models and we get the same conclusion as above, i.e., the local linear fit model with continuity at the knots is best.

6 Conclusions of the Analysis

By summarizing the content in the previous section, we find that preCD4 is significant but smoke is insignificant. To be specific, in linear mixed-effects models, if there were no polynomial function in the model, we might think that

smoking status had little effect on the percentage of CD4 cells. Also, in polynomial linear mixed-effects models, we may think that smoking status still does not have a great influence on the results, but for the second polynomial term, it is necessary to consider random intercepts and random coefficients with random slopes.

Then, in simple nonparametric models, we find that the cubic spline estimator is best among the three estimators. Also, in time-varying coefficient models, we find that there are little difference between those models.

Finally, in both piece-wise linear mixed-effects models and varying-coefficient mixed-effect models, we find that the local linear fit model with continuity at the knots is best choice.

7 Discussion of Advantages and Limitations

7.1 Linear mixed-effects model

Advantages: It is not necessary to assume that the observations are independent, and that there can be related observations in a unit.

Limitations: It needs to make more distribution assumptions. Also, it is often necessary to use approximation methods to estimate some parameters of the model so the conclusion relies on more assumptions, which increases the risk of

misspecifying the model and the resulting parameter estimation bias. Overall, they provide powerful and flexible tools for longitudinal data analysis.

7.2 Simple nonparametric model

Advantages: It can be applied in many situations and is more intuitive to understand. Also, it can be used for more types of data and smaller sample sizes.

Limitations: In parameter testing, it is usually neither powerful nor effective. In many cases, it is not as precise or accurate as parameter testing. It may lead to an incorrect decision due to insufficient accuracy of the test. Many of these tests do not take full advantage of the data in the analysis because they convert observed values into grades and groups.

7.3 Time-varying coefficient model

Advantages: At a given time, the coefficients of the different variables can be clearly seen, and we will see how these variables affect the response variable.

Limitations: As we described in the previous sections, it's not a good model, i.e., if we want to use this model, it would be difficult for us to tell which is better.

7.4 Piece-wise linear mixed-effects model

Advantages: It is easy to compute.

Limitations: It is hard to interpret the corresponding coefficients.

7.5 Varying-coefficient mixed-effects model

Advantages: It allows the linear coefficient to be a non-parametric curve with another variable.

Limitations: If it uses only a part of the sample, this may result in loss of information. If the complete sample is used, the model will take a long time to analyze.

8 Reference

[1] Wu and Tian. Nonparametric Models for Longitudinal Data: With Implementation in R, 2018, Chapman & Hall/CRC, New York

[2] Wu, Hulin, and Hua, Liang. “Backfitting Random Varying-Coefficient Models with Time-Dependent Smoothing Covariates”. Scandinavian Journal of Statistics, vol. 31, no. 1, 2004, pp. 3–19

[3] WU, C., 2020. NONPARAMETRIC MODELS FOR LONGITUDINAL DATA. [S.l.]: CRC PRESS

[4] Wikipedia-HIV. https://en.wikipedia.org/wiki/HIV/AIDS

[5] BMACS CD4 Dataset. https://www.rdocumentation.org/packages/npmlda/versions/1.0.0/topics/BMACS

Appendix (tables)

1 Head of variables

	ID	Time	Smoke	age	preCD4	CD4
1	1022	0.2	0	26.25	38	17
2	1022	0.8	0	26.25	38	30
3	1022	1.2	0	26.25	38	23
4	1022	1.6	0	26.25	38	15
5	1022	2.5	0	26.25	38	21
6	1022	3.0	0	26.25	38	12

2 AIC and BIC in simple linear mixed models with random intercept only

df	AIC	df	BIC
6	12561.80	6	12594.83
6	12521.08	6	12554.11
8	12523.26	8	12567.30

3 AIC and BIC in simple linear mixed models with random slope only

df	AIC	df	BIC
8	12164.18	8	12208.22
8	12114.56	8	12158.60
10	12115.68	10	12170.73

4 AIC and BIC in simple linear mixed models with both random intercept and slope

df	AIC	df	BIC
8	12164.18	8	12208.22
8	12114.56	8	12158.60
10	12115.68	10	12170.73

5 AIC and BIC in polynomial linear mixed models with random intercept only

df	AIC	df	BIC
8	12540.04	8	12584.07
8	12502.91	8	12546.95
11	12508.40	11	12568.96

6 AIC and BIC in polynomial linear mixed models with random slope only

df	AIC	df	BIC
10	12140.67	10	12195.72
10	12095.78	10	12150.83
13	12100.46	13	12172.02

7 AIC and BIC in polynomial linear mixed models with both random intercept and slope

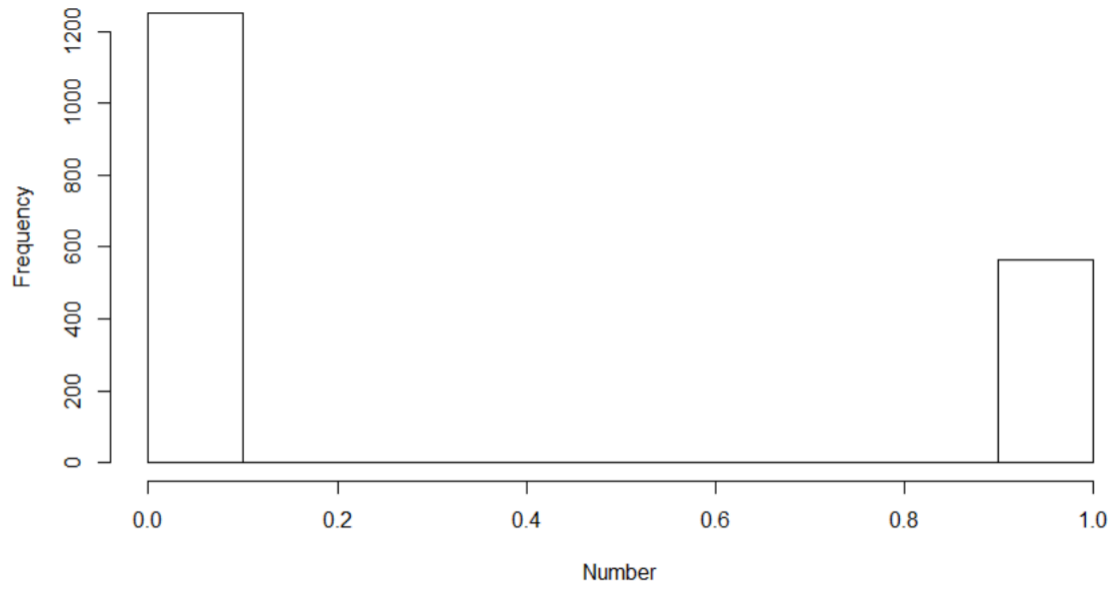
df	AIC	df	BIC
10	12140.67	10	12195.72
10	12095.78	10	12150.83
13	12100.46	13	12172.02

8 AIC and BIC in polynomial linear mixed models with random intercept, slope and coefficient

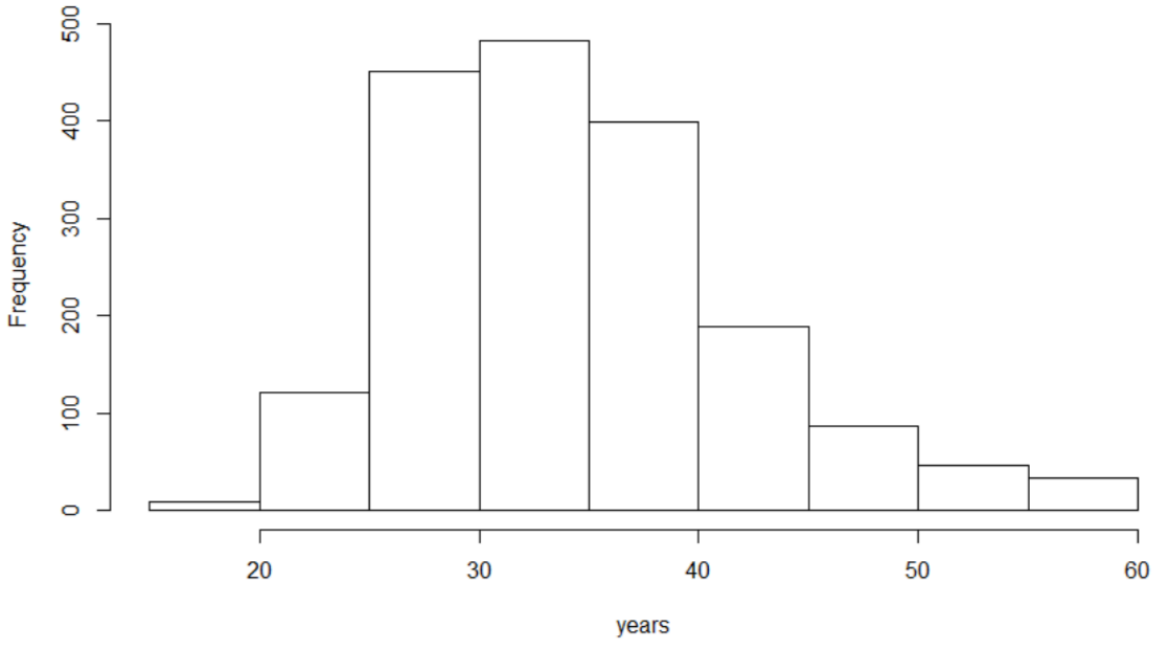
df	AIC	df	BIC
13	12113.93	13	12185.49
13	12068.41	13	12139.97
16	12073.42	16	12161.50

Appendix (figures and code)

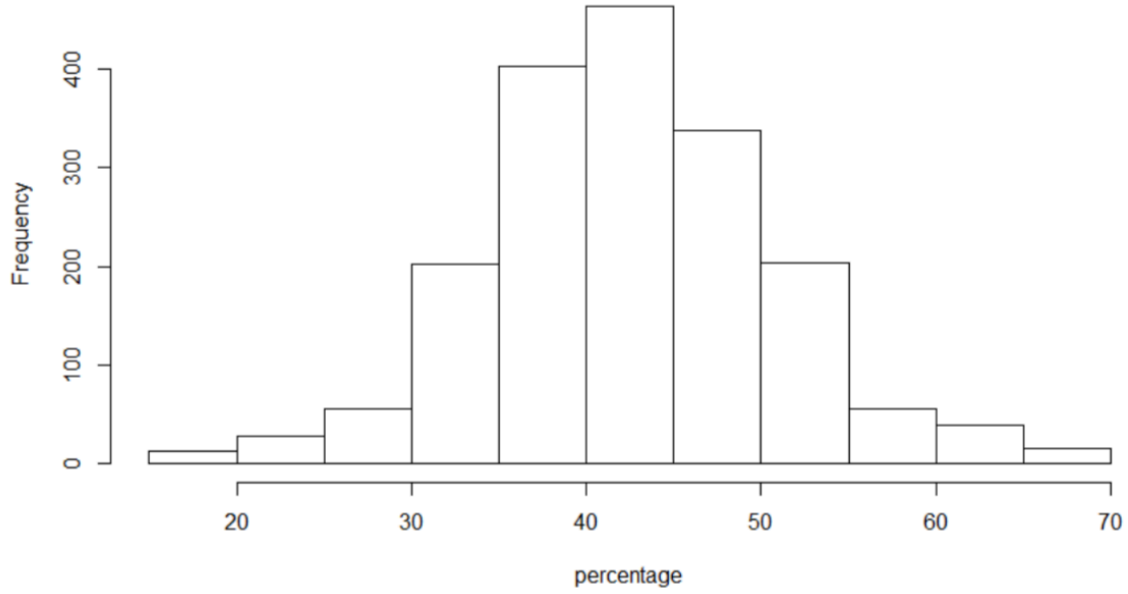
1 Distribution of smoke



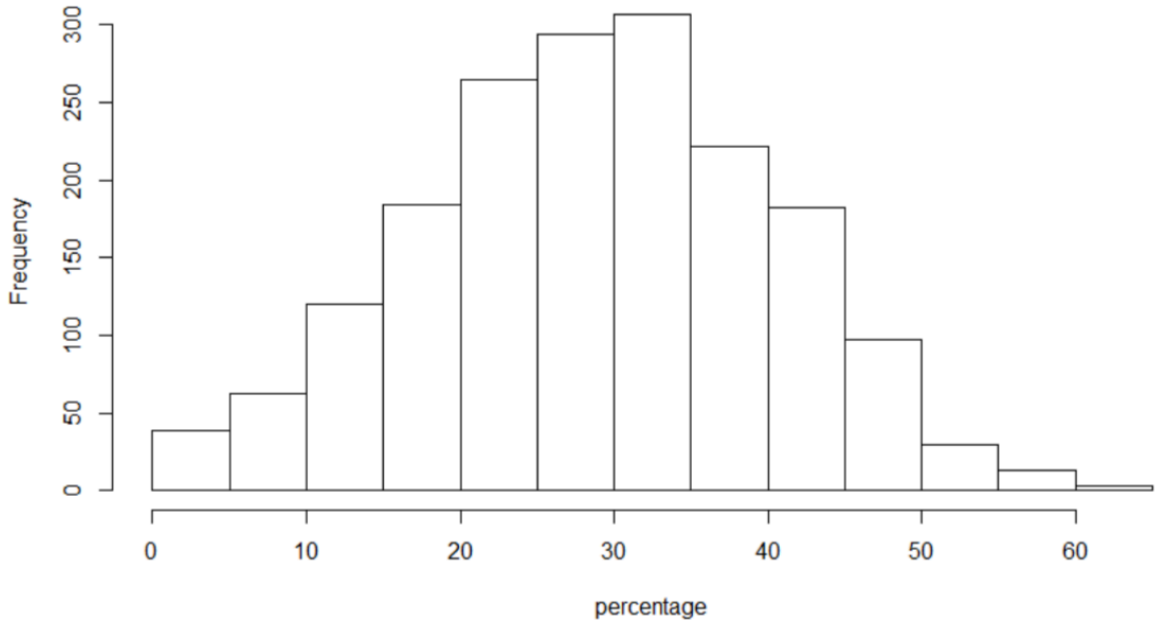
2 Distribution of age



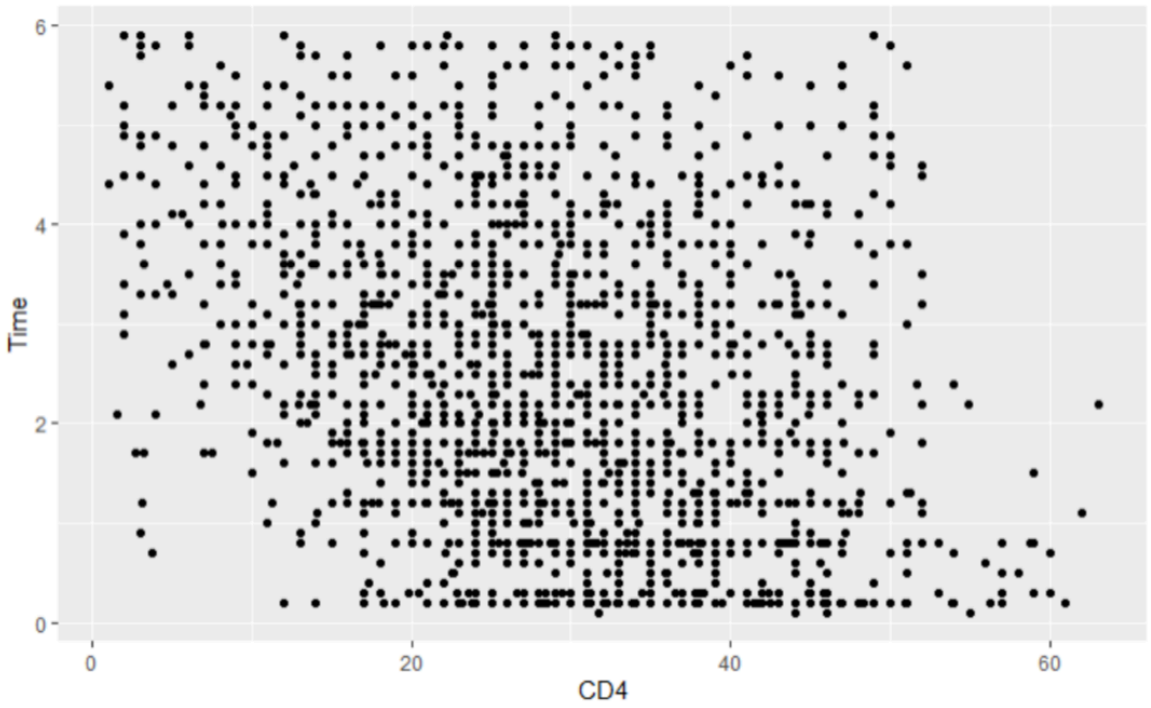
3 Distribution of preCD4



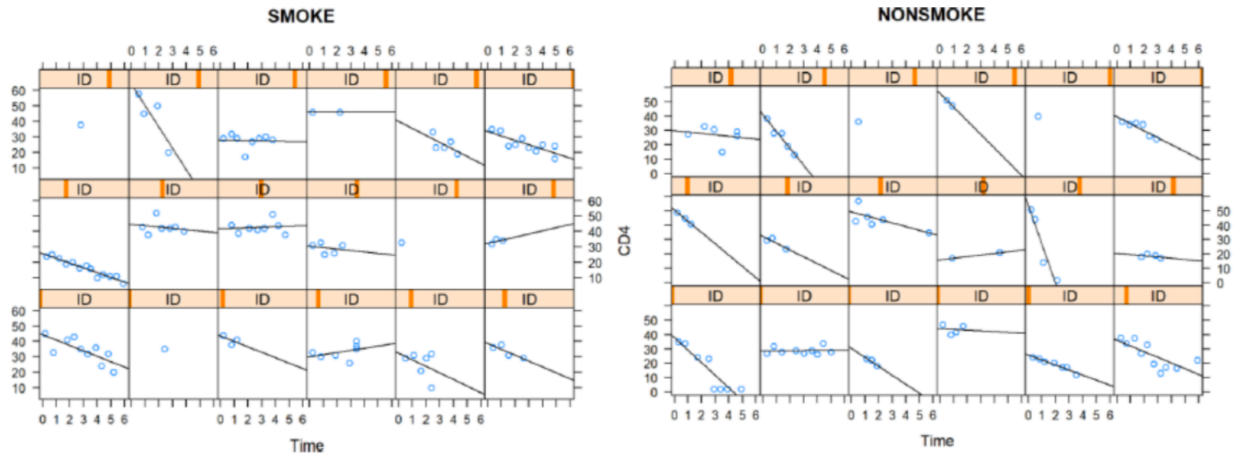
4 Distribution of CD4



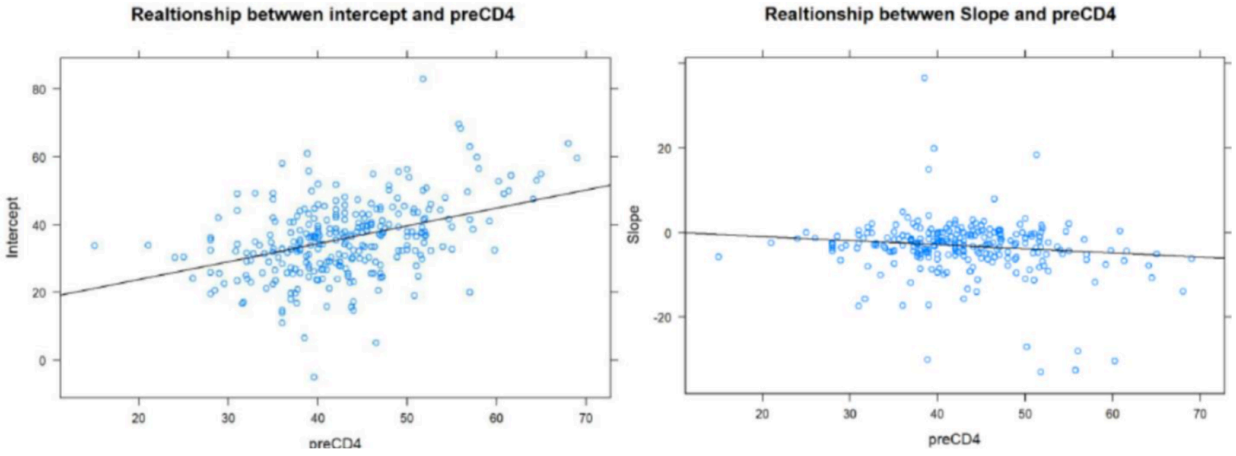
5 Relationship between time and CD4



6 Relationship between time and CD4 for patients in smoke and non-smoke group



7 Relationship between preCD4 and intercept or slope



8 anova for m2 and m4 and anova for m2 and m6

```
> anova(fit2,fit4)
refitting model(s) with ML (instead of REML)
Data: x
Models:
fit2: CD4 ~ 1 + Time * preCD4 + (1 | ID)
fit4: CD4 ~ 1 + Time * Smoke + (Time | ID)
    Df   AIC   BIC  logLik deviance  Chisq Chi Df Pr(>Chisq)
fit2  6 12508 12541 -6248.2    12496
fit4  8 12166 12210 -6075.1    12150 346.21      2  < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> anova(fit2,fit6)
refitting model(s) with ML (instead of REML)
Data: x
Models:
fit2: CD4 ~ 1 + Time * preCD4 + (1 | ID)
fit6: CD4 ~ 1 + Time * preCD4 + Time * Smoke + (Time | ID)
    Df   AIC   BIC  logLik deviance  Chisq Chi Df Pr(>Chisq)
fit2  6 12508 12541 -6248.2    12496
fit6 10 12108 12163 -6044.1    12088 408.22      4  < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

9 anova for lm2 and lm5, lm5 and lm8, and lm5 and lm11

```
> anova(fit.mod2,fit.mod5)
```

```
refitting model(s) with ML (instead of REML)
Data: x
Models:
fit.mod2: CD4 ~ 1 + Time * preCD4 + I(Time^2) * preCD4 + (1 | ID)
fit.mod5: CD4 ~ 1 + Time * preCD4 + I(Time^2) * preCD4 + (Time | ID)
       Df   AIC   BIC  logLik deviance  Chisq Chi Df Pr(>Chisq)
fit.mod2  8 12478 12523 -6231.3    12462
fit.mod5 10 12074 12129 -6026.9    12054 408.71      2  < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> anova(fit.mod5,fit.mod8)
refitting model(s) with ML (instead of REML)
Data: x
Models:
fit.mod5: CD4 ~ 1 + Time * preCD4 + I(Time^2) * preCD4 + (Time | ID)
fit.mod8: CD4 ~ 1 + Time * preCD4 + I(Time^2) * preCD4 + (1 + Time | ID)
       Df   AIC   BIC  logLik deviance Chisq Chi Df Pr(>Chisq)
fit.mod5 10 12074 12129 -6026.9    12054
fit.mod8 10 12074 12129 -6026.9    12054     0      0         1
> anova(fit.mod5,fit.mod11)
refitting model(s) with ML (instead of REML)
Data: x
Models:
fit.mod5: CD4 ~ 1 + Time * preCD4 + I(Time^2) * preCD4 + (Time | ID)
fit.mod11: CD4 ~ 1 + Time * preCD4 + I(Time^2) * preCD4 + (1 + Time + I(Time^2) |
fit.mod11:     ID)
       Df   AIC   BIC  logLik deviance  Chisq Chi Df Pr(>Chisq)
```

fit.mod5 10 12074 12129 -6026.9   12054

fit.mod11 13 12047 12119 -6010.7   12021 32.525   3 4.056e-07 ***

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

10 AIC for m2 and lm11(best 2 models in linear mixed models)

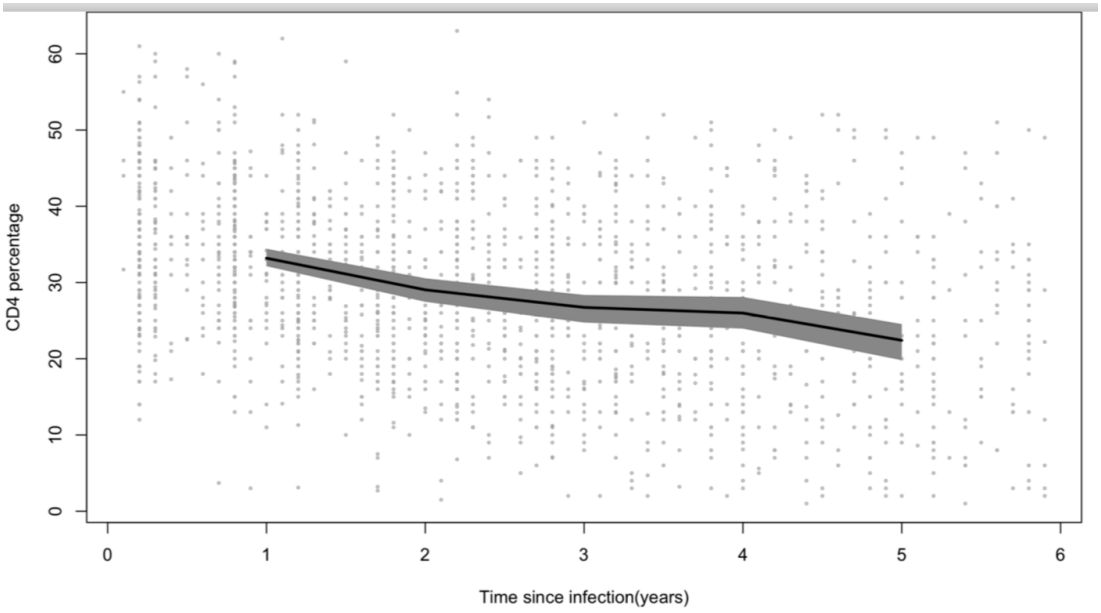
> AIC(fit2,fit.mod11)

        df   AIC

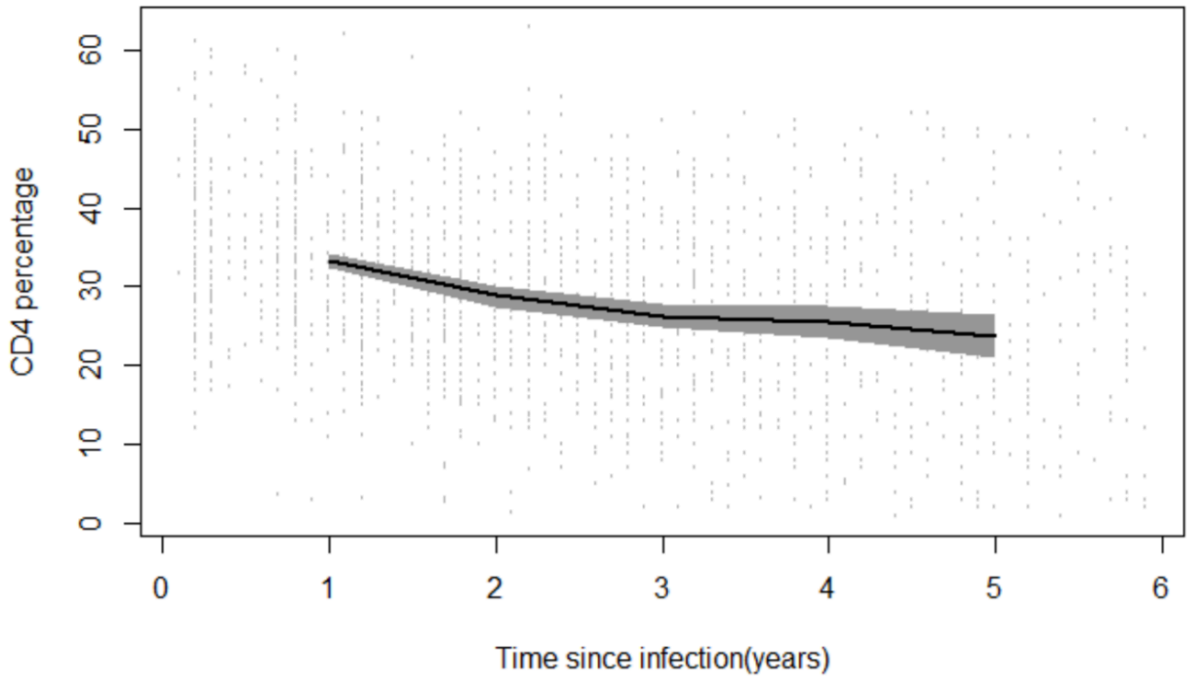
fit2   6 12521.08

fit.mod11 13 12068.41

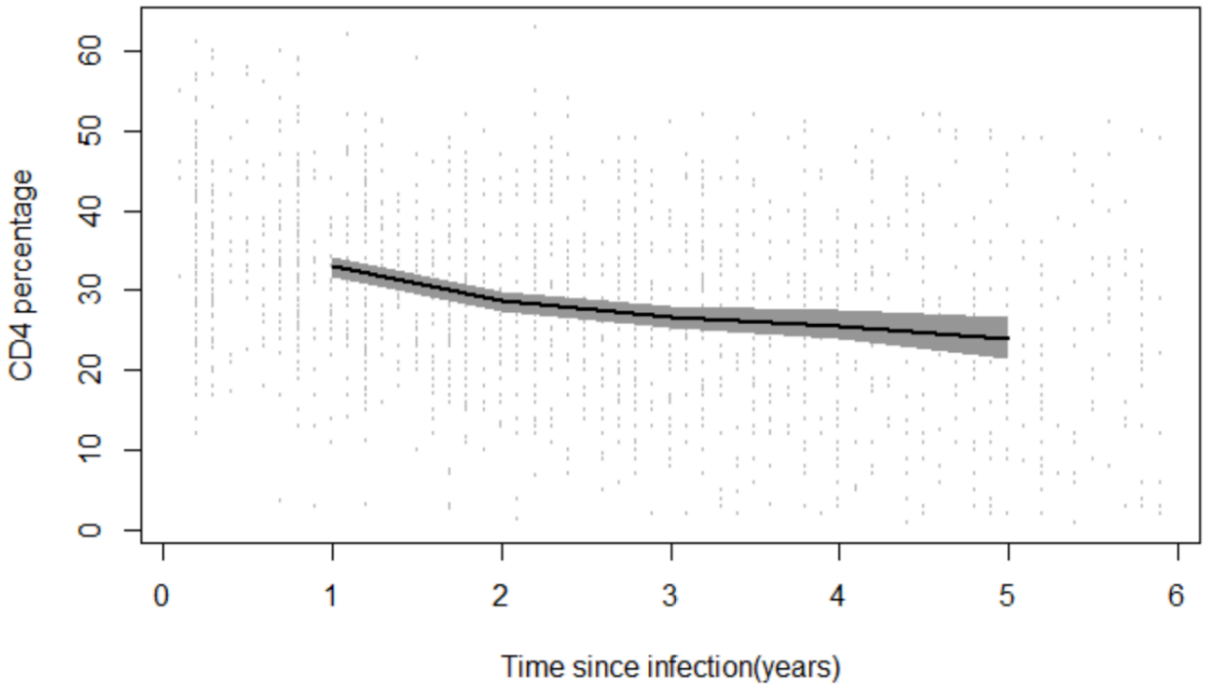
11 kernel estimator using the Uniform (-1/2, 1/2) kernel and bandwidth h = 0.5



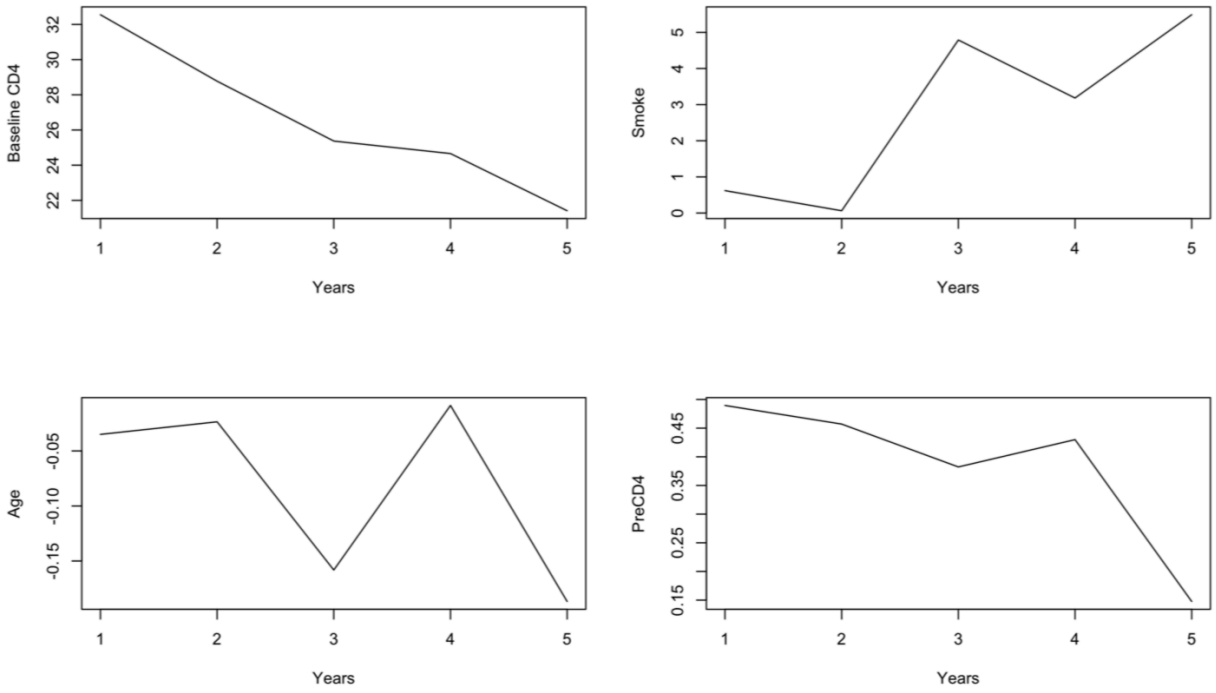
12 kernel estimator using the Uniform (-1/2, 1/2) kernel and bandwidth h = 1



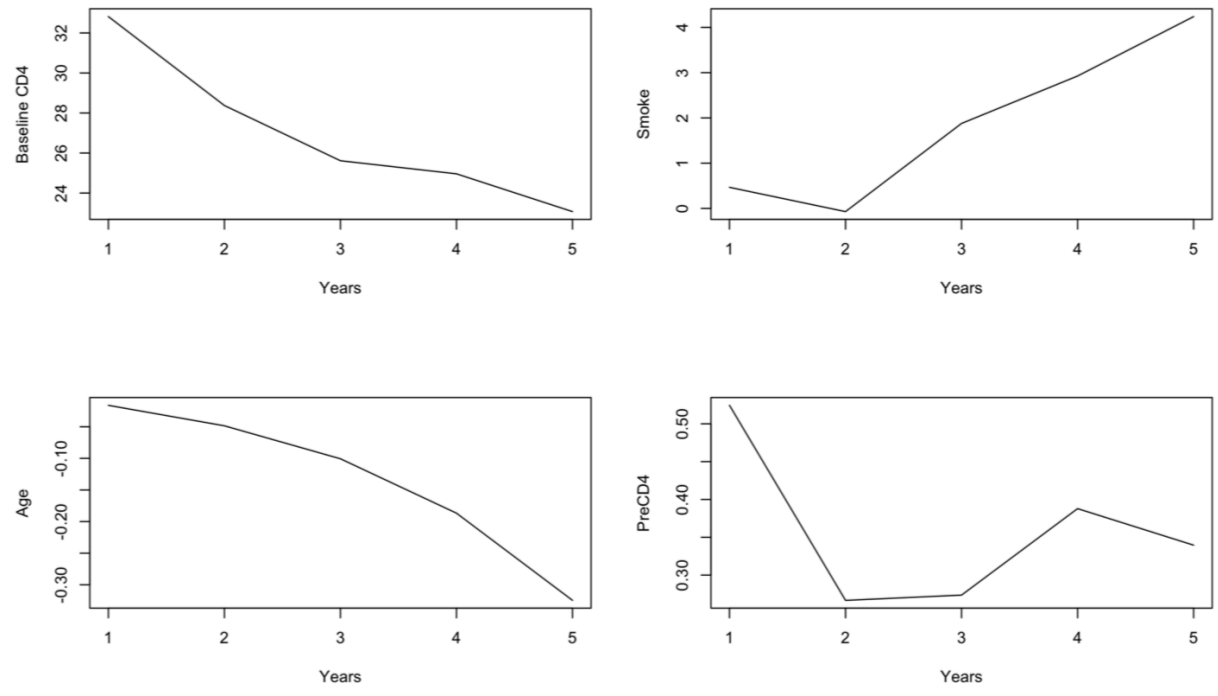
13 spline estimator using the cubic spline with two interior knots



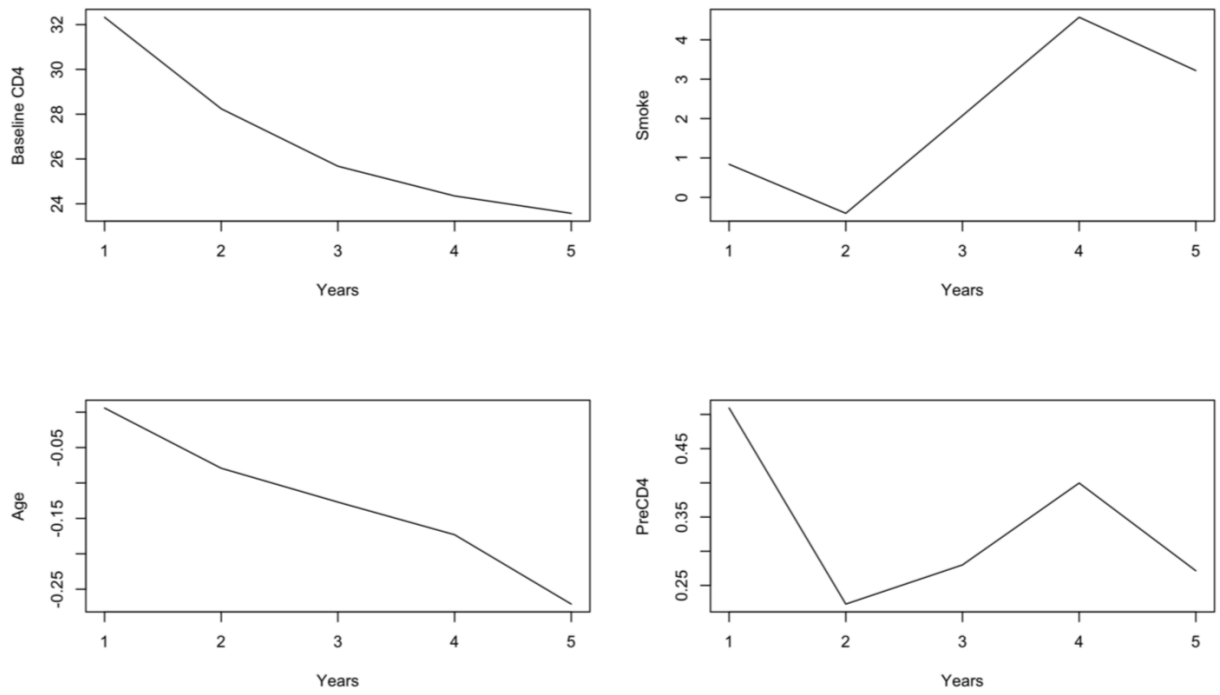
14 kernel estimator using the Uniform (-1/2, 1/2) kernel and bandwidth h = 0.5



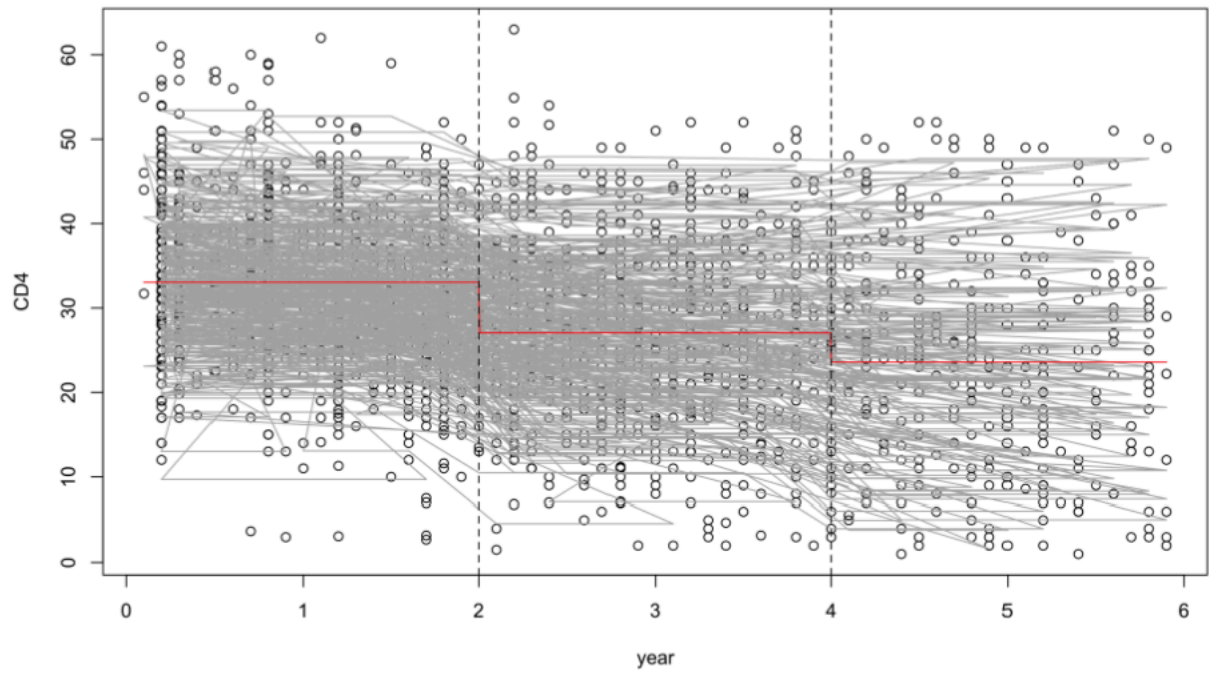
15 kernel estimator using the Uniform (-1/2, 1/2) kernel and bandwidth h = 1



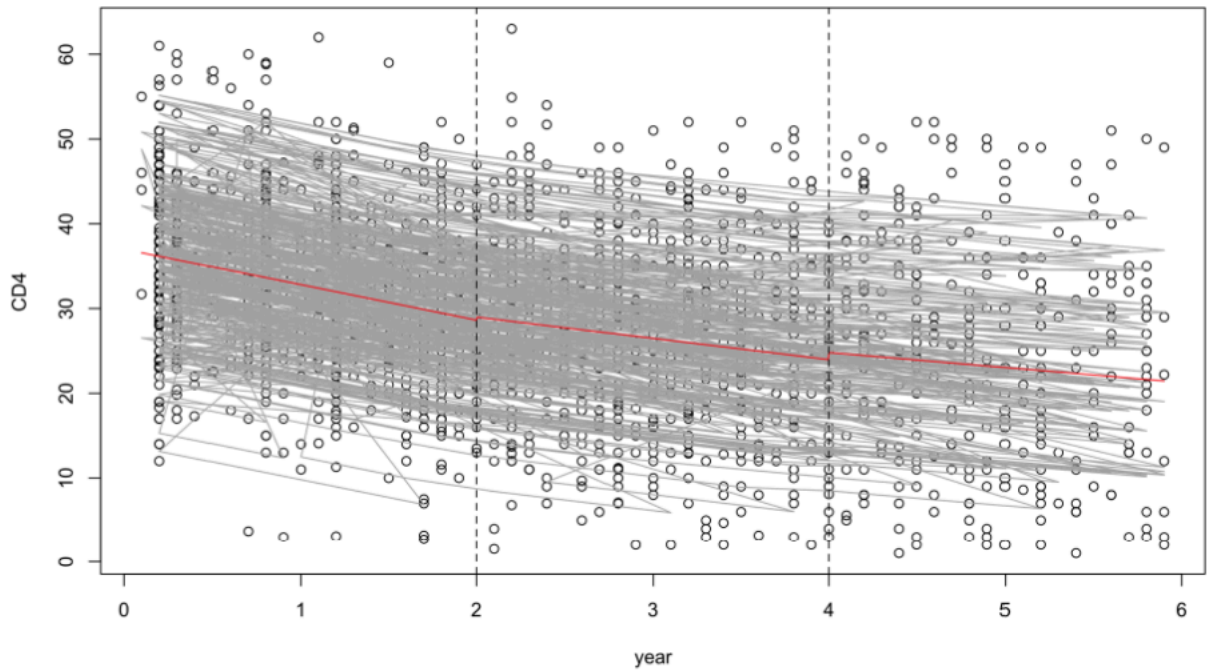
16 plot of spline estimator using the cubic spline with two interior knots



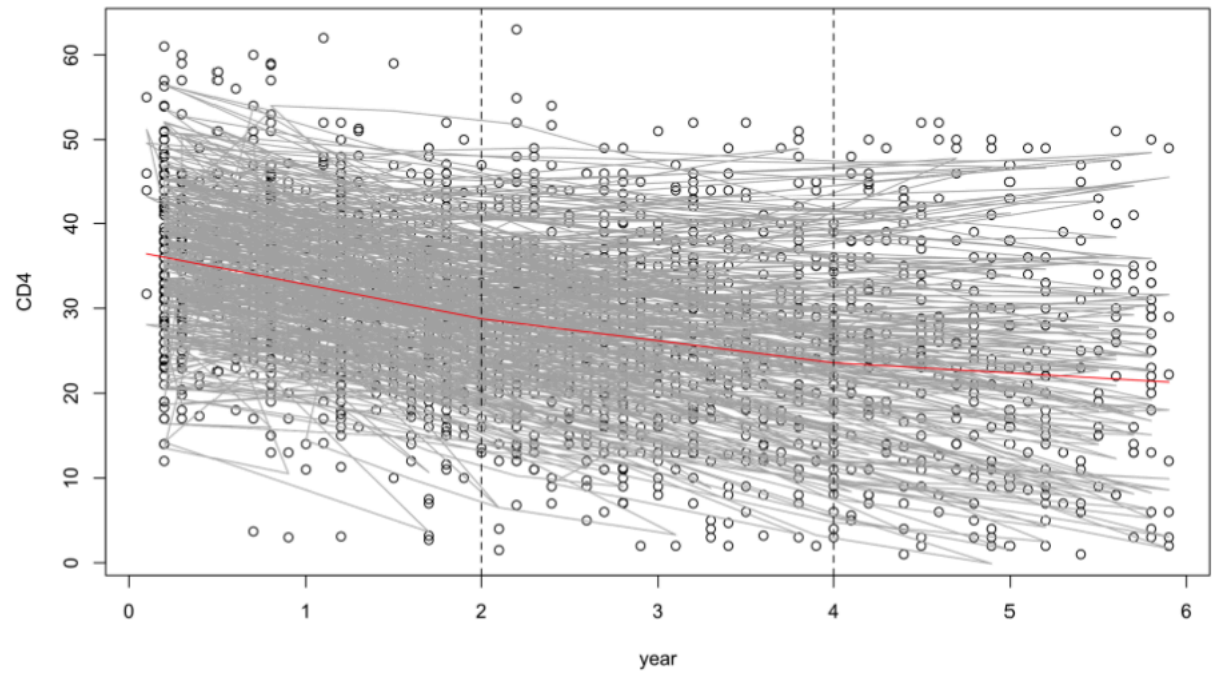
17 local constant fit using the basic functions B1(t) = 1[t<2], B2(t) = 1[2<t<4] and B3(t) = 1[4<t]



18 local linear fit without assuming continuity at the knots



19 local linear fit with assuming continuity at the knots



20 ANOVA among simple piece-wise linear mixed-effects models

> anova(fit1,fit2)

refitting model(s) with ML (instead of REML)

Data: BMACS

Models:

fit2: CD4 ~ 0 + basis2(Time) + (1 | ID)

fit1: CD4 ~ 0 + basis1(Time) + (1 + basis1(Time) | ID)

Df AIC BIC logLik deviance Chisq Chi Df Pr(>Chisq)

fit2 8 12535 12579 -6259.6 12519

fit1 14 12445 12522 -6208.5 12417 102.23

6 < 2.2e-16 ***

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

> anova(fit1,fit3)

refitting model(s) with ML (instead of REML)

Data: BMACS

Models:

fit1: CD4 ~ 0 + basis1(Time) + (1 + basis1(Time) | ID)

fit3: CD4 ~ basis3(Time) + (1 + basis3(Time) | ID)

Df AIC BIC logLik deviance Chisq Chi Df Pr(>Chisq)

fit1 14 12445 12522 -6208.5 12417

fit3 20 12116 12226 -6038.0 12076 341

6 < 2.2e-16 ***

—

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

> (cbind(CI.lower,CI.upper))

CI.lower CI.upper

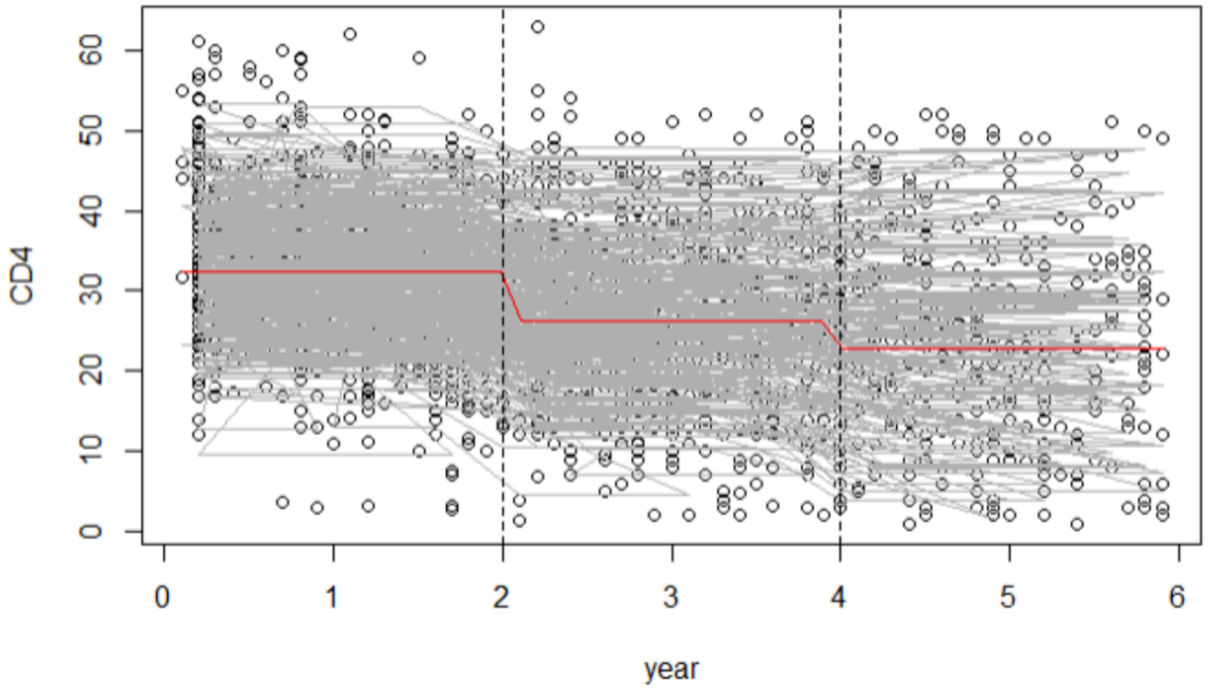
(Intercept) 35.7114147 38.241289

basis3(Time)2 -4.8733722 -3.282402

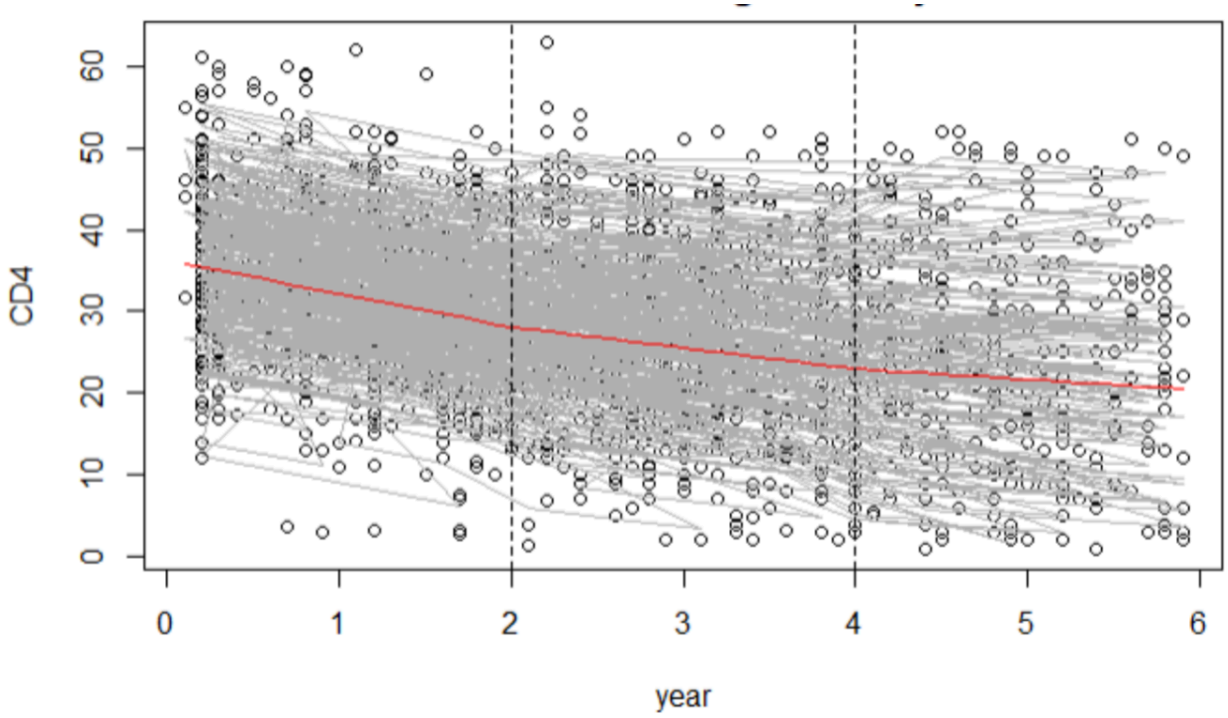
basis3(Time)3 0.5172552 2.764431

basis3(Time)4 -0.2218978 2.382884

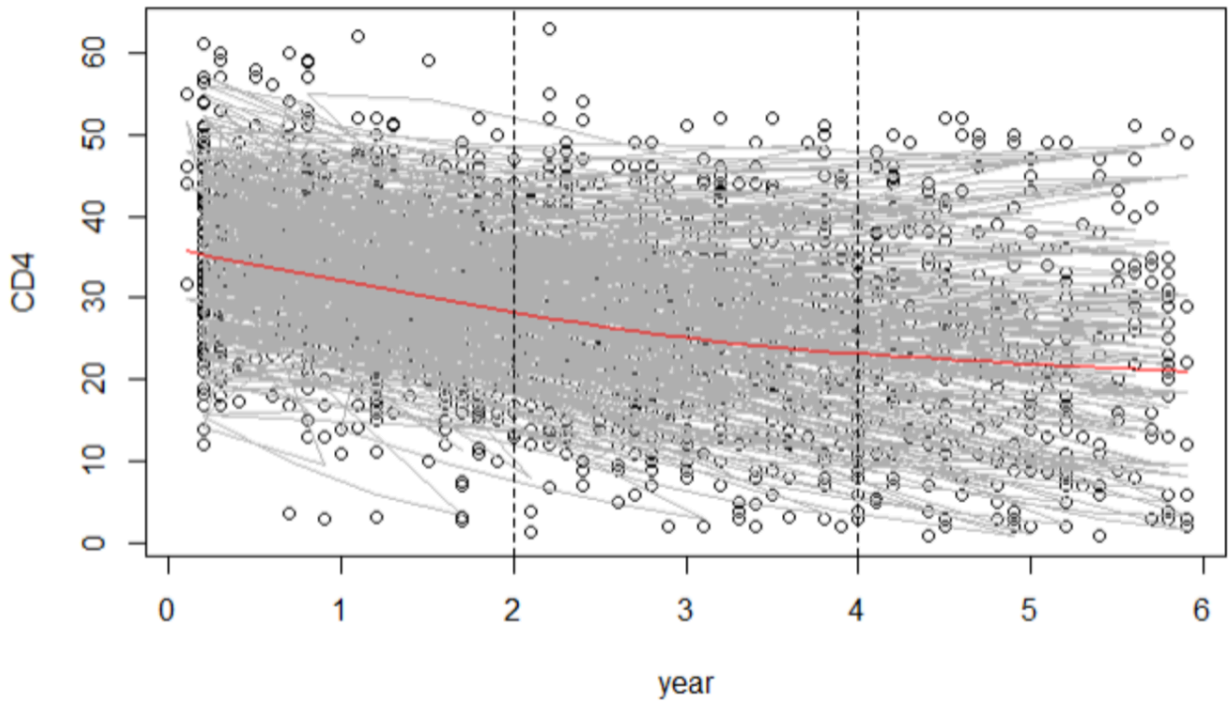
21 local constant fit using the basic functions B1(t) = 1[t<2], B2(t) = 1[2<t<4] and B3(t) = 1[4<t]



22 local linear fit without assuming continuity at the knots



23 local linear fit with assuming continuity at the knots



24 ANOVA among varying-coefficient mixed-effects models

> anova(fit4,fit5)

refitting model(s) with ML (instead of REML)

Data: BMACS

Models:

fit4: CD4 ~ 0 + basis1(Time) + basis4(Time, Smoke) + (0 + basis1(Time) |

fit4: ID)

fit5: CD4 ~ 0 + basis2(Time) + basis5(Time, Smoke) + (1 | ID)

Df AIC BIC logLik deviance Chisq Chi Df Pr(>Chisq)

fit4 13 12437 12508 -6205.4 12411

fit5 14 12535 12612 -6253.6 12507 0

> anova(fit4,fit6)

refitting model(s) with ML (instead of REML)

Data: BMACS

Models:

fit4: CD4 ~ 0 + basis1(Time) + basis4(Time, Smoke) + (0 + basis1(Time) |

fit4: ID)

fit6: CD4 ~ 0 + basis3(Time) + basis6(Time, Smoke) + (0 + basis3(Time) |

fit6: ID)

Df AIC BIC logLik deviance Chisq Chi Df Pr(>Chisq)

fit4 13 12437 12508 -6205.4 12411

fit6 19 12109 12214 -6035.5 12071 339.65

6 < 2.2e-16 ***

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

> merBoot1 = bootMer(x = fit6, FUN = fixef, nsim = 100)

> CI.lower = apply(merBoot1$t, 2, function(x) as.numeric(quantile(x, probs=.025,

na.rm=TRUE)))

> CI.upper = apply(merBoot1$t, 2, function(x) as.numeric(quantile(x, probs=.975,

na.rm=TRUE)))

> (cbind(CI.lower,CI.upper))

	CI.lower	CI.upper
basis3(Time)1	34.4492756	37.7571864
basis3(Time)2	-4.7985149	-3.0889339
basis3(Time)3	0.1145417	2.3330854
basis3(Time)4	0.6453930	3.4112040

basis6(Time, Smoke)1 -0.4295782 5.0035174

basis6(Time, Smoke)2 -2.0669538 1.4157211

basis6(Time, Smoke)3 -1.4992028 2.7555428

basis6(Time, Smoke)4 -4.3915045 0.4275039