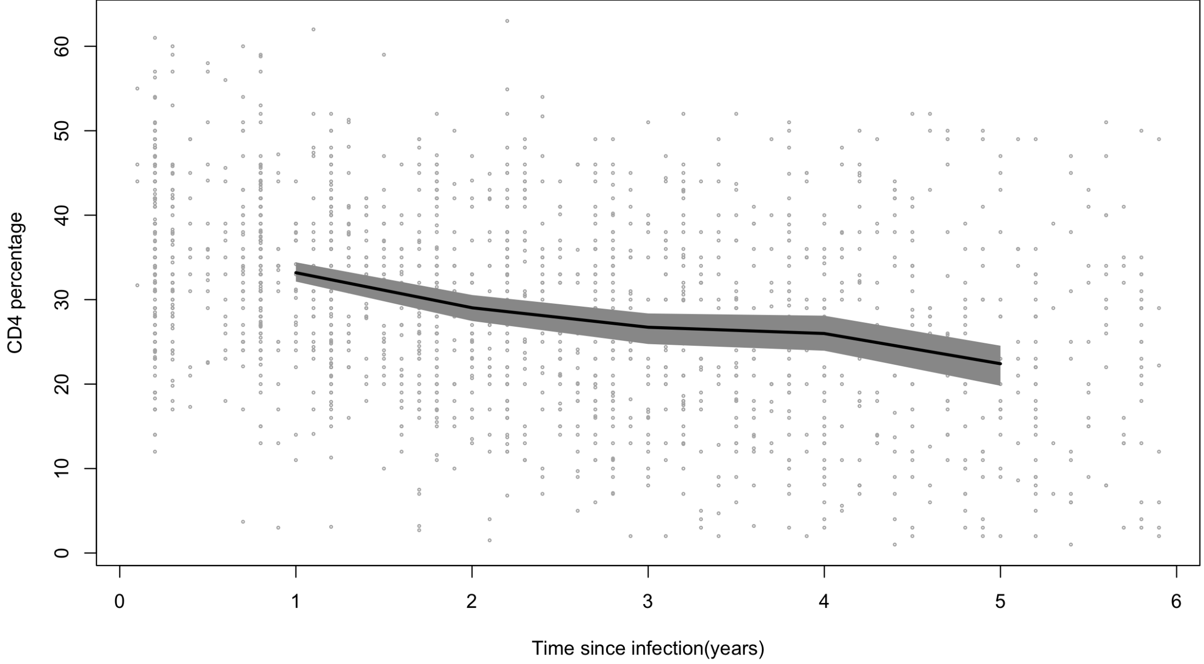
1

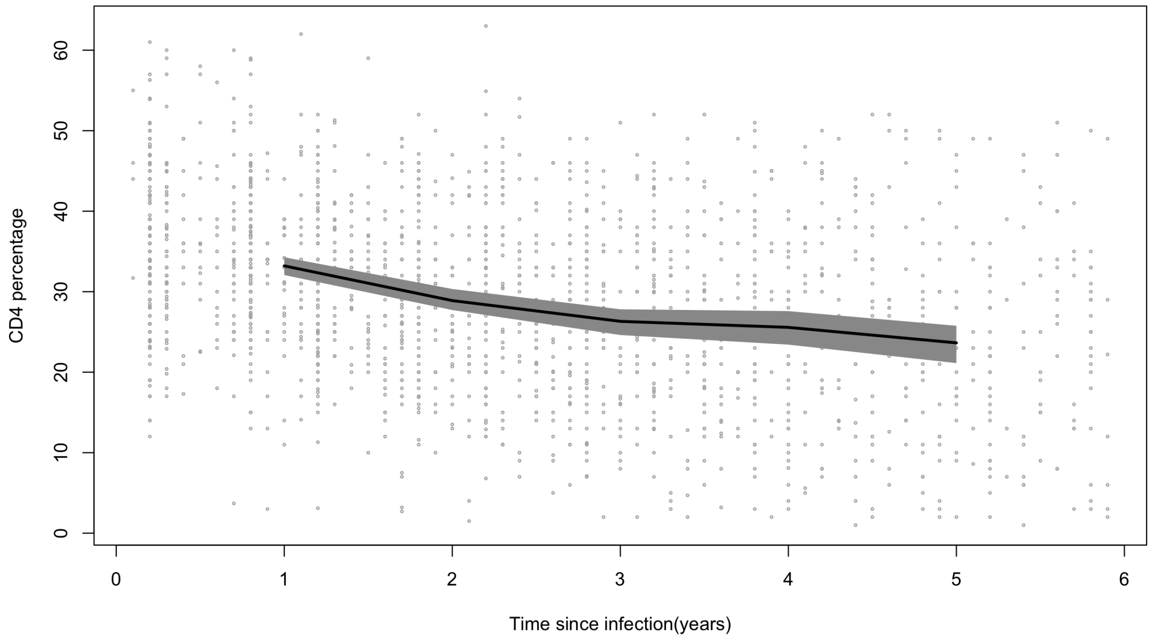
1.1

```
> set.seed(123)
> library(npmlda)
> library(splines)
> library(plyr)
> BMACS=read.csv(“/Users/Aaron/Desktop/BMACS.csv")
> head(BMACS)
    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
> uni = function(bw,data){
+   t.int = c(1,2,3,4,5)
+   mu.t = c()
+   i=1
+   for (time in t.int) {
+     t1 = time-bw/2
+     t2 = time+bw/2
+     cd4 = data[data[,2]>=t1 & data[,2]<=t2,6]
+     mu.t[i] = sum(cd4)/length(cd4)
+     i = i+1
+   }
+   return(mu.t) }
> (mu1 = uni(0.5,BMACS))
[1] 33.19116 29.04009 26.73218 25.99058 22.41250
> IDlist = unique(BMACS$ID)
> nID = length(IDlist)
> Bootsample = function(){
+   resample.ID = sample(IDlist ,nID ,replace=T)
+   do.call("rbind", lapply(1:nID, function(i) subset(BMACS, ID==resample.ID[i]))) }
> Time.int = c(1,2,3,4,5)
> Boot.Fit = replicate(200, uni(0.5,Bootsample()))
> UpperCI = apply(Boot.Fit, 1, quantile,.975)
> LowerCI = apply(Boot.Fit, 1, quantile,.025)
> (cbind(Time.int,LowerCI,UpperCI))
     Time.int  LowerCI  UpperCI
[1,]        1 32.03151 34.42934
[2,]        2 27.54351 30.55911
[3,]        3 24.99483 28.57383
[4,]        4 24.05633 28.45995
[5,]        5 20.38278 25.14932
> plot(CD4 ~ Time, data = BMACS, xlab = "Time since infection(years)", ylab = "CD4
percentage", cex=0.3, col="gray70", main="")
> polygon(c(Time.int[1], Time.int, rev(Time.int)),c(LowerCI[1], UpperCI, rev(LowerCI)),
col="gray60", border=NA)
> lines(Time.int, mu1, lwd=2.5, col=1)
```



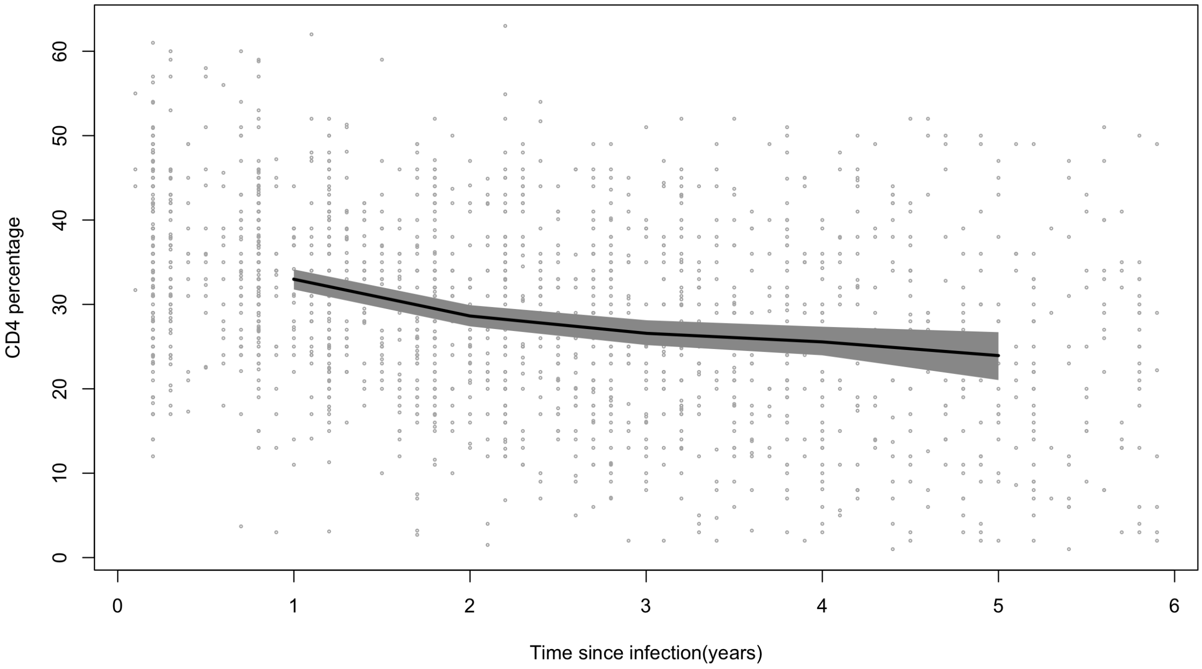
1.2

```
> (mu2 = uni(1,BMACS))
[1] 33.20507 28.89281 26.31755 25.56854 23.64278
> Time.int = c(1,2,3,4,5)
> Boot.Fit = replicate(200, uni(1,Bootsample()))
> UpperCI = apply(Boot.Fit, 1, quantile,.975)
> LowerCI = apply(Boot.Fit, 1, quantile,.025)
> (cbind(Time.int,LowerCI,UpperCI))
     Time.int  LowerCI  UpperCI
[1,]        1 32.06918 34.28025
[2,]        2 27.71972 30.33535
[3,]        3 24.61847 27.81492
[4,]        4 23.42256 27.58786
[5,]        5 21.12816 25.76475
> plot(CD4 ~ Time, data = BMACS, xlab = "Time since infection(years)", ylab = "CD4
percentage", cex=0.3, col="gray70", main="")
> polygon(c(Time.int[1], Time.int, rev(Time.int)),c(LowerCI[1], UpperCI, rev(LowerCI)),
col="gray60", border=NA)
> lines(Time.int, mu2, lwd=2.5, col=1)
```



1.3

```
> Time.int = c(1:5)
> sp = function(data0){
+   fit = lm(CD4~bs(Time,knots=c(2,4)),data=data0)
+   mu = predict(fit,newdata = list(Time=Time.int))
+   return(mu)
+ }
> (mu3 = sp(BMACS))
       1        2        3        4        5 
32.98256 28.62889 26.57958 25.56030 23.93812 
> Boot.Fit = replicate(200, sp(Bootsample()))
> UpperCI = apply(Boot.Fit, 1, quantile,.975)
> LowerCI = apply(Boot.Fit, 1, quantile,.025)
> (cbind(Time.int,LowerCI,UpperCI))
  Time.int  LowerCI  UpperCI
1        1 31.77323 34.14500
2        2 27.40136 29.91279
3        3 25.18831 28.13032
4        4 23.97039 27.36127
5        5 21.03888 26.71151
> plot(CD4 ~ Time, data = BMACS, xlab = "Time since infection(years)", ylab = "CD4 percentage", cex=0.3, col="gray70", main="")
> polygon(c(Time.int[1], Time.int, rev(Time.int)),c(LowerCI[1], UpperCI, rev(LowerCI)), col="gray60", border=NA)
> lines(Time.int, mu3, lwd=2.5, col=1)
```



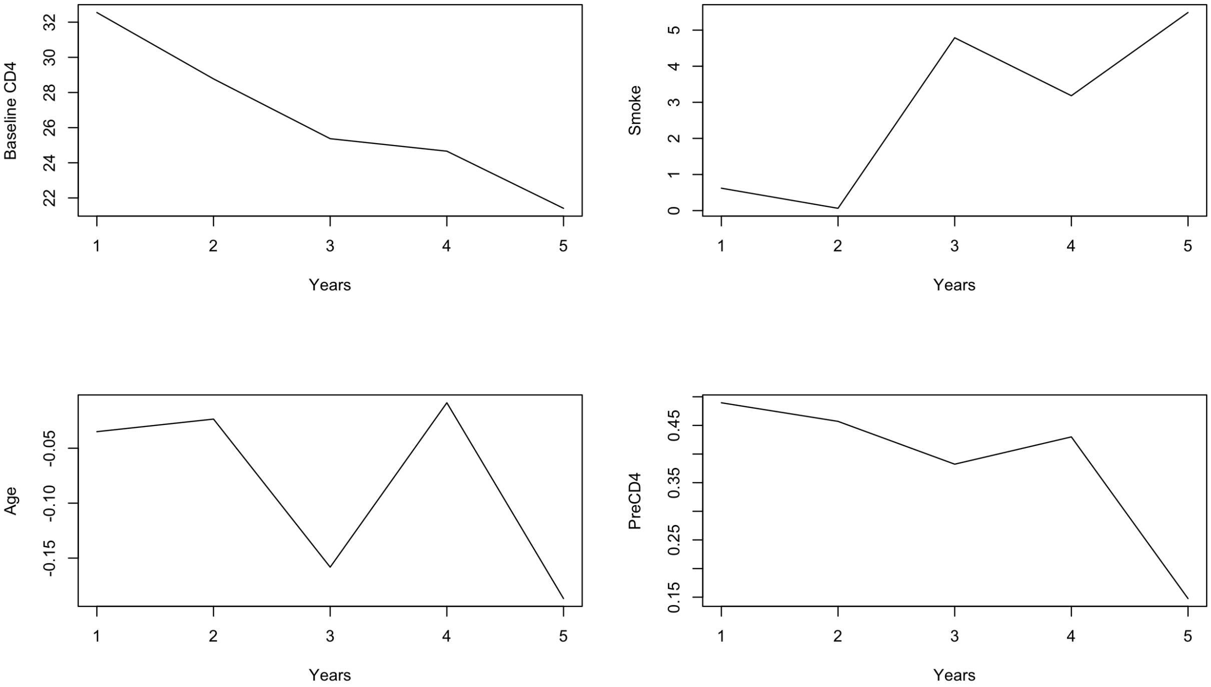
1.4

It shows that the curve returned by the cubic spline estimator with two knots is smoother, i.e., the spline is better than the uniform estimators and the curve is smoother when the bandwidth is larger for the uniform estimator. Therefore, compared with the other two uniform estimates, we can conclude that the cubic spline estimator is a better choice.

2

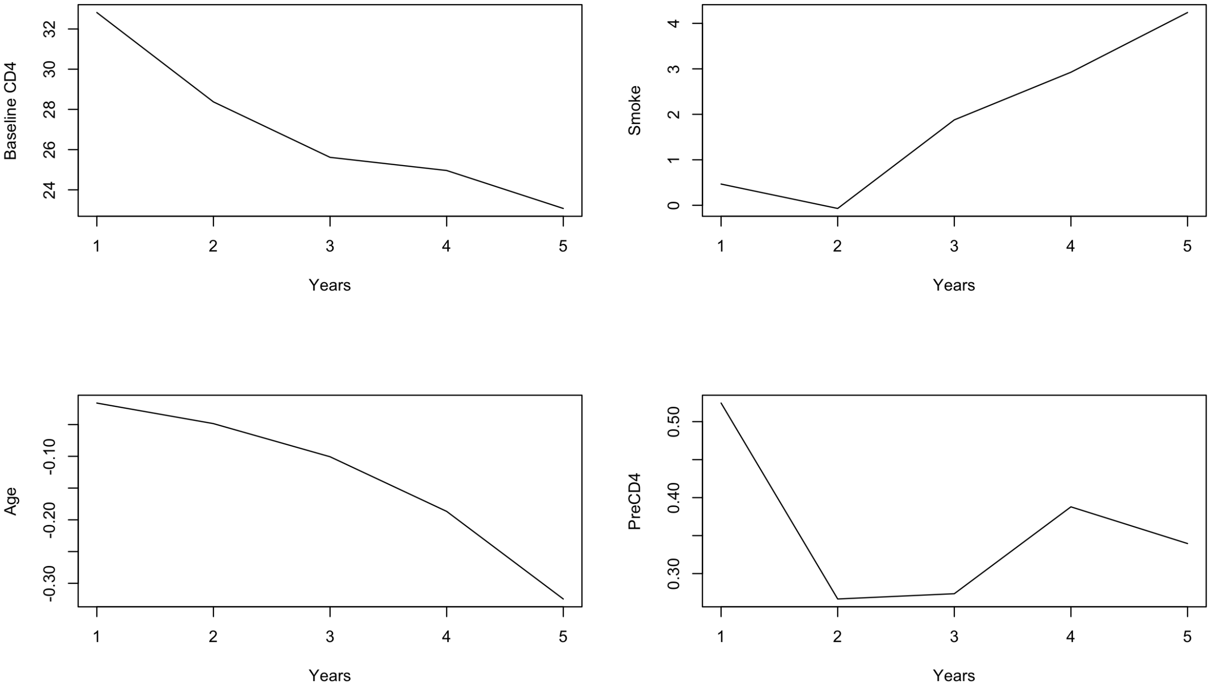
2.1

```
> N = nrow(BMACS)
> CountID = count(BMACS$ID)
> colnames(CountID) = c("Subject ID","ni")
> n = nrow(CountID)
> N = nrow(BMACS)
> Pdata = c()
> for ( i in 1 :n){Pdata = rbind(Pdata,BMACS[which(BMACS$ID==CountID[i,1]),c(4,5)][1,]) }
> mp = apply(Pdata,2,mean)
> BMACS2 = cbind(BMACS[,c(1:3)],BMACS[,4]-rep(mp[1],N),BMACS[,5]-rep(mp
[2],N),BMACS[,6])
> colnames(BMACS2) = c("ID","time","X1","X2","X3","CD4")
> K = function(x="predict value",T="tij",h="bandwidth") I(abs(T-x)<=0.5*h)
> est = function(t){
+ C = function(i){
+   tID = BMACS2[which(BMACS2$ID==CountID[i,1]),]
+   ni = nrow(tID)
+   Di = as.matrix(cbind(rep(1,ni),tID[,c(3:5)]))
+   KI = function(j) K(t,tID$time[j],h)
+   Ki = diag(ni)*(sapply(1:ni,KI))
+   return(cbind( t(Di)%*%Ki%*%Di/ni,t(Di)%*%Ki%*%tID[,6]/ni)) }
+ c = matrix(0,nrow = 4,ncol=5)
+ for (i in 1:n){
+   Tei = C(i)
+   c = c+Tei}
+ return(solve(c[,1:4])%*%c[,5]) }
> h = 0.5
> Est1 = sapply(1:5,est)
> row.names(Est1) = c("Intercept","Beta1","Beta2","Beta3")
> colnames(Est1) = c("time = 1","time = 2","time = 3","time = 4","time = 5")
> show(Est1)
             time = 1    time = 2   time = 3     time = 4   time = 5
Intercept 32.54966425 28.77323961 25.3712960 24.660732991 21.4105214
Beta1      0.62068423  0.06325858  4.7881459  3.183092911  5.4880810
Beta2     -0.03493861 -0.02350184 -0.1582259 -0.008641498 -0.1867900
Beta3      0.48956480  0.45705743  0.3822764  0.429940262  0.1476084
> par(mfrow=c(2,2))
> plot(1:5,Est1[1,],xlab = "Years",ylab = "Baseline CD4",type = "l")
> plot(1:5,Est1[2,],xlab = "Years",ylab = "Smoke",type = "l")
> plot(1:5,Est1[3,],xlab = "Years",ylab = "Age",type = "l")
> plot(1:5,Est1[4,],xlab = "Years",ylab = "PreCD4",type = "l")
```



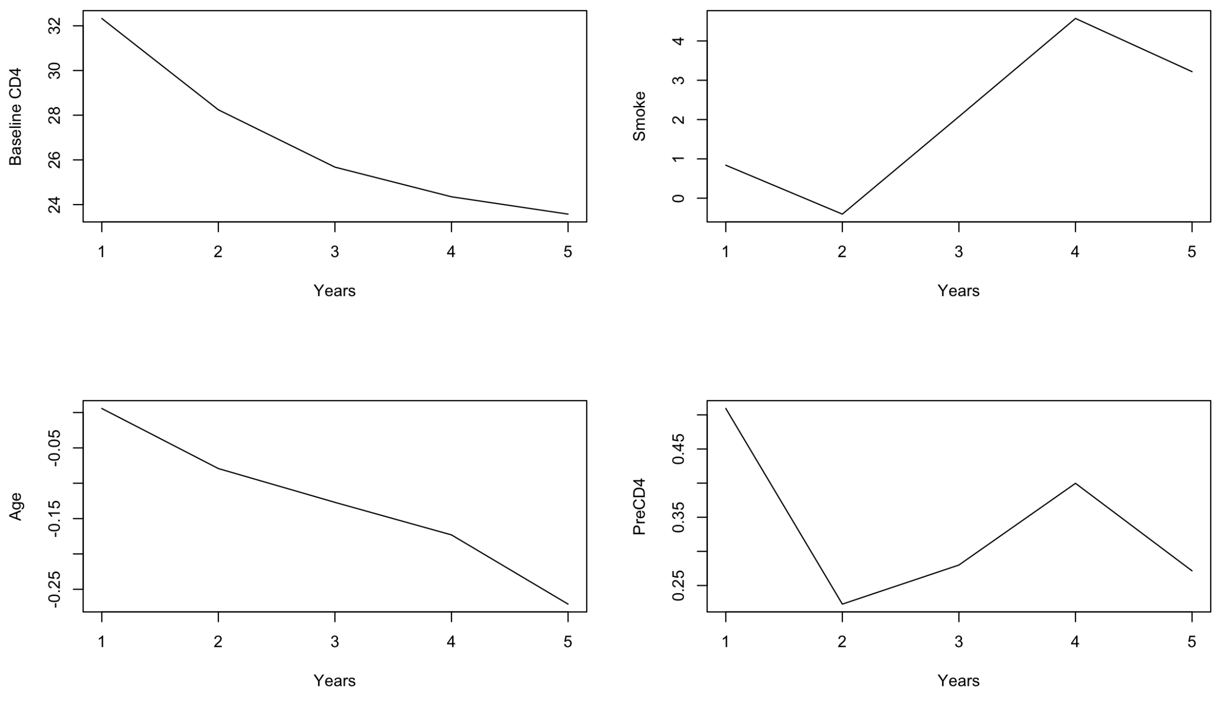
2.2

```
> h=1
> Est2 = sapply(1:5,est)
> row.names(Est2) = c("Intercept","Beta1","Beta2","Beta3")
> colnames(Est2) = c("time = 1","time = 2","time = 3","time = 4","time = 5")
> show(Est2)
             time = 1     time = 2    time = 3    time = 4    time = 5
Intercept 32.81838803 28.37389869 25.6151723 24.9613416 23.0728220
Beta1      0.46686688 -0.07006122  1.8778366  2.9258006  4.2387369
Beta2     -0.01610827 -0.04849689 -0.1006723 -0.1867172 -0.3246527
Beta3      0.52446962  0.26659033  0.2735383  0.3878733  0.3395516
> par(mfrow=c(2,2))
> plot(1:5,Est2[1,],xlab = "Years",ylab = "Baseline CD4",type = "l")
> plot(1:5,Est2[2,],xlab = "Years",ylab = "Smoke",type = "l")
> plot(1:5,Est2[3,],xlab = "Years",ylab = "Age",type = "l")
> plot(1:5,Est2[4,],xlab = "Years",ylab = "PreCD4",type = “l")
```



2.3

```
> basis = function(t){ cbind(t,t^2,t^3,I(t-2>0)*(t-2)^3,I(t-4>0)*(t-4)^3)}
> B = function(t){M = matrix(0,nrow = 4,ncol = 24)
+   for (i in 1:4){
+     M[i,c((6*(i-1)+1):(6*(i-1)+6))] = cbind(1,basis(t)) }
+  return(M) }
> Spi = function(i){
+   tid = BMACS2[which(BMACS2$ID==CountID[i,1]),]
+   ni = nrow(tid)
+   Uj = matrix(0,nrow = ni,ncol = 24)
+   for(j in 1:ni){
+     Bj = B(tid$time[j])
+     Uj[j,] = as.matrix(cbind(1,tid[j,c(3,4,5)]))%*%Bj }
+   wi = diag(ni)/ni
+   return(cbind(t(Uj)%*%wi%*%Uj,t(Uj)%*%wi%*%tid[,6])) }
> Sp = matrix(0,nrow = 24,ncol=25)
> for (i in 1:n){ Ti = Spi(i)
+ Sp = Sp+Ti}
> est = solve(Sp[,1:24])%*%Sp[,25]
> Es = function(t) B(t)%*%est
> Est3 = sapply(1:5,Es)
> row.names(Est3) = c("Intercept","beta1","beta2","beta3")
> colnames(Est2) = c("time = 1","time = 2","time = 3","time = 4","time = 5")
> show(Est3)
                 [,1]        [,2]       [,3]       [,4]       [,5]
Intercept 32.328774132 28.24227070 25.6766272 24.3491435 23.5732682
beta1      0.839638590 -0.40525654  2.0756556  4.5722078  3.2188731
beta2      0.005793324 -0.07921368 -0.1270061 -0.1730616 -0.2709906
beta3      0.509410351  0.22273508  0.2800239  0.3997323  0.2714513
> par(mfrow=c(2,2))
> plot(1:5,Est3[1,],xlab = "Years",ylab = "Baseline CD4",type = "l")
> plot(1:5,Est3[2,],xlab = "Years",ylab = "Smoke",type = "l")
> plot(1:5,Est3[3,],xlab = "Years",ylab = "Age",type = "l")
> plot(1:5,Est3[4,],xlab = "Years",ylab = "PreCD4",type = “l")
```



2.4

We compared uniform estimate coefficients with bandwidth between 0.5 and 1 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.

3

In assignment 1, we used different parametric models to fit the data, which usually required many preconditions or assumptions to ensure good model results. Also, we can easily fit the model to produce the results as there are fewer constraints. However, we can see that under nonparametric models it is usually difficult to apply and it may reduce the work efficiency of the model due to some restrictions.

As a result, both models have advantages and disadvantages and if we know that the parametric model is a good fit for our data we will prefer to apply this model rather than the non-parametric model.

4

In assignment 2, we can see that the uniform kernel estimator usually has a closer bootstrap confidence interval with larger bandwidth, but spline functions may be more efficient than the uniform kernel estimator although it is difficult to find the difference. It means that 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.