Longitudinal changes in a population of interest are often heterogeneous and may be influenced by a combination of baseline factors. The longitudinal tree (that is, regression tree with longitudinal data) can be very helpful to identify and characterize the sub-groups with distinct longitudinal profile in a heterogenous population. This blog presents the capabilities of the R package LongCART for constructing longitudinal tree according to the LongCART algorithm (Kundu and Harezlak 2019). In addition, this packages can also be used to formally evaluate whether any particular baseline covariate affects the longitudinal profile via parameter instability test. In this blog, construction of longitudinal tree is illlustrated with an R dataset in step by step approach and the results are explained.

Installing and Loading LongCART package

```
R> install.packages("LongCART")
R> library(LongCART)
```

Get the example dataset

The ACTG175 dataset in speff2trial package contains longitudinal observation of CD4 counts from clinical trial in HIV patients. This dataset is in “wide” format, and, we need to convert it to first “long” format.

```
R> library(speff2trial)
R> data("ACTG175", package = "speff2trial")
R> adata<- reshape(data=ACTG175[,!(names(ACTG175) %in% c("cd80", "cd820"))],
+ varying=c("cd40", "cd420", "cd496"), v.names="cd4",
+ idvar="pidnum", direction="long", times=c(0, 20, 96))
R> adata<- adata[order(adata$pidnum, adata$time),]
```

Longtudinal model of interest

Since the count data including CD4 counts are often log transformed before modeling, a simple longitudinal model for CD4 counts would be:
log(CD4 countit) = beta0 + beta1*t + bi + epsilonit

Does the fixed parameters of above longitdinal vary with the level of baseline covariate?

Categorical baseline covariate

Suppose we want to evaluate whether any of the fixed model parameters changes with the levels of any baseline categorical partitioning variable, say, gender.

```
R> adata$Y=ifelse(!is.na(adata$cd4),log(adata$cd4+1), NA)
R> StabCat(data=adata, patid="pidnum", fixed=Y~time, splitvar="gender")
Stability Test for Categorical grouping variable
Test.statistic=0.297, p-value=0.862
```

The p-value is 0.862 which indicates that we don’t have any evidence that fixed parameters vary with the levels of gender.

Continuous baseline covariate

Now suppose we are interested to evaluate whether any of the fixed model parameters changes with the levels of any baseline continuous partitioning variable, say, wtkg.

```
R> StabCont(data=adata, patid="pidnum", fixed=Y~time, splitvar="wtkg")
Stability Test for Continuous grouping variable
Test.statistic=1.004 1.945, Adj. p-value=0.265 0.002
```

The result returns two two p-values – the first p-value correspond to parameter instability test of beta0 and the second ones correspond to beta1 .

Constructing tree for longitudinal profile

The ACTG175 dataset contains several baseline variables including gender, hemo (presence of hemophilia), homo (homosexual activity), drugs (history of intravenous drug use), oprior (prior non-zidovudine antiretroviral therapy), z30 (zidovudine use 30 days prior to treatment initiation), zprior (zidovudine use prior to treatment initiation), race, str2 (antiretroviral history), treat (treatment indicator), offtrt (indicator of off-

treatment before 96 weeks), age, wtkg (weight) and karnof (Karnofsky score). We can construct longitudinal tree to identify the sub-groups defined by these baseline variables such that the individuals within the given sub-groups are homogeneous with respect to longitudinal
profile of CD4 counts but the longitudinal profiles among the sub-groups are heterogenous.

```
R> gvars=c("age", "gender", "wtkg", "hemo", "homo", "drugs",
+ "karnof", "oprior", "z30", "zprior", "race",
+ "str2", "symptom", "treat", "offtrt", "strat")
R> tgvars=c(1, 0, 1, 0, 0, 0,
+ 1, 0, 0, 0, 0,
+ 0, 0, 0, 0, 0)
R> out.tree<- LongCART(data=adata, patid="pidnum", fixed=Y~time,
+ gvars=gvars, tgvars=tgvars, alpha=0.05,
+ minsplit=100, minbucket=50, coef.digits=3)
```

All the baseline variables are listed in gvars argument. The gvars argument is accompanied with the tgvars argument which indicates type of the partitioning variables (0=categorical or 1=continuous). Note that the LongCART() function currently can handle the categorical variables with numerical levels only. For nominal variables, please create the corresponding numerically coded dummy variable(s).

Now let’s view the tree results

```
R> out.tree$Treeout
ID n yval var index p (Instability) loglik improve Terminal
1 1 2139 5.841-0.003time offtrt 1.00 0.000 -4208 595 FALSE
2 2 1363 5.887-0.002time treat 1.00 0.000 -2258 90 FALSE
3 4 316 5.883-0.004time str2 1.00 0.005 -577 64 FALSE
4 8 125 5.948-0.002time symptom NA 1.000 -176 NA TRUE
5 9 191 5.84-0.005time symptom NA 0.842 -378 NA TRUE
6 5 1047 5.888-0.001time wtkg 68.49 0.008 -1645 210 FALSE
7 10 319 5.846-0.002time karnof NA 0.260 -701 NA TRUE
8 11 728 5.907-0.001time age NA 0.117 -849 NA TRUE
9 3 776 5.781-0.007time karnof 100.00 0.000 -1663 33 FALSE
10 6 360 5.768-0.009time wtkg NA 0.395 -772 NA TRUE
11 7 416 5.795-0.005time z30 1.00 0.014 -883 44 FALSE
12 14 218 5.848-0.003time treat NA 0.383 -425 NA TRUE
13 15 198 5.738-0.007time age NA 0.994 -444 NA TRUE
```

In the above output, each row corresponds to single node including the 7 terminal nodes identified by TERMINAL=TRUE. Now let’s visualize the tree results

```
R> par(xpd = TRUE)
R> plot(out.tree, compress = TRUE)
R> text(out.tree, use.n = TRUE)
```

The resultant tree is as follows:
ACTG175tree