Project 2

Clare Clingain

January 20, 2018

Data

The New York City Department of Education (NYC DOE) published datasets containing the 2013-2017 ELA state exam results for schools in New York City. Scores of 1 and 2 are viewed as inadequate, while 3 and 4 are passing. 5th grade is a crucial time point in which students will either be promoted to middle school, or held back in elementary school. For this analysis, 696 schools' 5th grade state exam results of scores 3 or 4 were examined longitudinally. The outcome variable, Y, is the percentage of students scoring a 3 or a 4 in each public elementary school.

In all of New York City, there are 810 public elementary schools. However, data was not reported for all schools due to FERPA regulations which stipulate that if a student or students can be easily identified from their data, it cannot be publicly released. For this dataset, cases like these mainly consisted of public schools in which only a few students took the 5th grade state exam, so performance could easily be matched to a student when paired with race, economic status, English Language Learner (ELL) status, and disability. Therefore, the dataset I used was reduced to 696 schools.
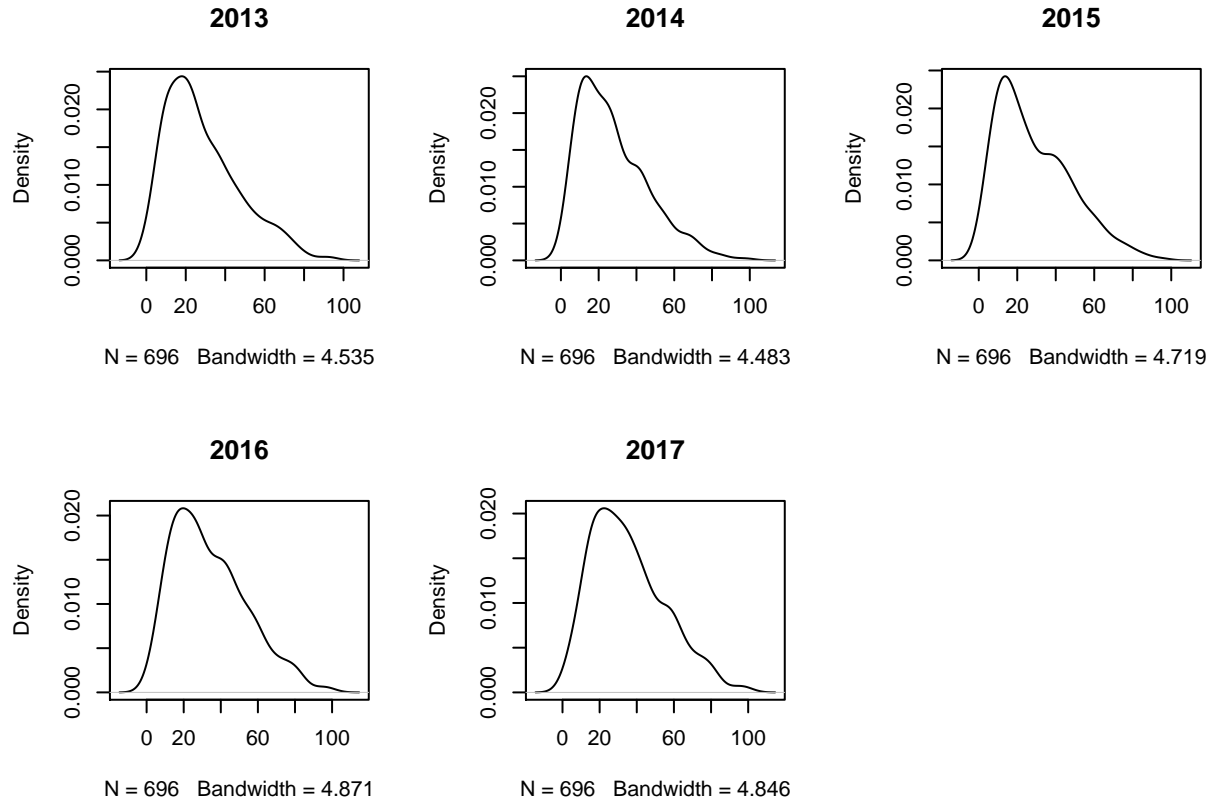
Analysis

Pre-Processing Transformations

Initially, I had to clean up the dataset a lot! All of the years were listed in the same column, with all the respective scores listed in the same column, as well. Therefore, I arranged the data so that each score was associated in its own column for each year (i.e., "One.2013" or "Four.2017"). Also, the percents weren't importing correctly due to symbols used that did not translate to proper decimals, so I calculated the correct percents, bound by 0 and 100.

My five variables of interest that make up my feature set did not need any transformation due to relatively symmetric distributions (see below). Since all five variables are percentages, they all share the same minimum and maximum, so re-scaling wasn't needed either.

```
#slight skew, but reasonably symmetric
par(mfrow=c(2,3))
plot(density(ela.na.5$Three.Four.Per.2013), main = "2013")
plot(density(ela.na.5$Three.Four.Per.2014), main = "2014")
plot(density(ela.na.5$Three.Four.Per.2015), main = "2015")
plot(density(ela.na.5$Three.Four.Per.2016), main = "2016")
plot(density(ela.na.5$Three.Four.Per.2017), main = "2017")
```

2013 **2014** **2015**

Density

N = 696   Bandwidth = 4.535   N = 696   Bandwidth = 4.483   N = 696   Bandwidth = 4.719

2016 **2017**

Density

N = 696   Bandwidth = 4.871   N = 696   Bandwidth = 4.846

Clustering Method

Using the lcmm package in R, I conducted an approximation of Nagin clustering on my data. All clustering was done on the raw components, not the principal components. In Nagin clustering, one can add time-invariant and time-variant covariates in an effort to explain the means. Additionally, each cluster can be assigned any order polynomial that best describes the structure of that cluster. However, in the lcmm package, all clusters have to receive the same order polynomail (i.e., all given quadratic term), which limits my analysis. Therefore, it is possible that a higher-order model would be best if multiple clusters have significant quadratic terms even if its BIC is lower.

The number of clusters to use was determined by BIC in the Mclust function from the Mclust package in R. Mclust automatically reports the model that has the highest BIC, which is defined as $LL - 0.5k * ln(N)$, in which N is the number of schools. Given the cluster membership g, $Y_i(t) = \mu_{gi}(t) + \epsilon_i(t)$, where $\mu_{gi}(t) = b_0 + b_1(t) + b_2(t)^2 + ... + b_k(t)^k$ and the error terms are independent across time and have identical variance. Thus, Y_i varies as a function of time. However, Mclust does not allow for as complex error term structures as Nagin clustering in Stata.

Given the number of suggested clusters from Mclust, I ran six different models. The first set were fixed models, one with only linear terms, the other with quadratic terms. Next, I ran linear and quadratic models, but this time with random intercept. Random intercept was included as random=~1. Finally, I added linear and quadratic random slope models, with Time as the variant (random=~Time).

Upon chosing the best model, I added the time-invariant risk factor of Brooklyn. This allowed me to determine whether group membership is conditional on a school's geographical location, namely Brooklyn. I decided to use Brooklyn due to its larger student population compared to the other outerboroughs and Manhattan. Out

of all five boroughs, Brooklyn was the most represented in the dataset (221 schools), followed by Queens (173 schools), the Bronx (130 schools), Manhattan (121 schools), and Staten Island (42 schools).

Finally, I made a brief comparison to a non model-based k-means solution since it is important to consider other methods and solutions in clustering because of the assumptions made regarding the data.

Mclust results

First, I graphed my data pre-colored/labeled.

```r
par(mfrow=c(1,1))
matplot(t(ela.na.5[,c(8,12,16,20,24)]),type='l',col=1,pch=1,xlab='Year',
        ylab='% Receiving 3 or 4 on NYS ELA exam', main = "% 5th graders in Public
        Schools Receiving Score of 3 or 4", cex.lab = .75, cex.main = .75)
```
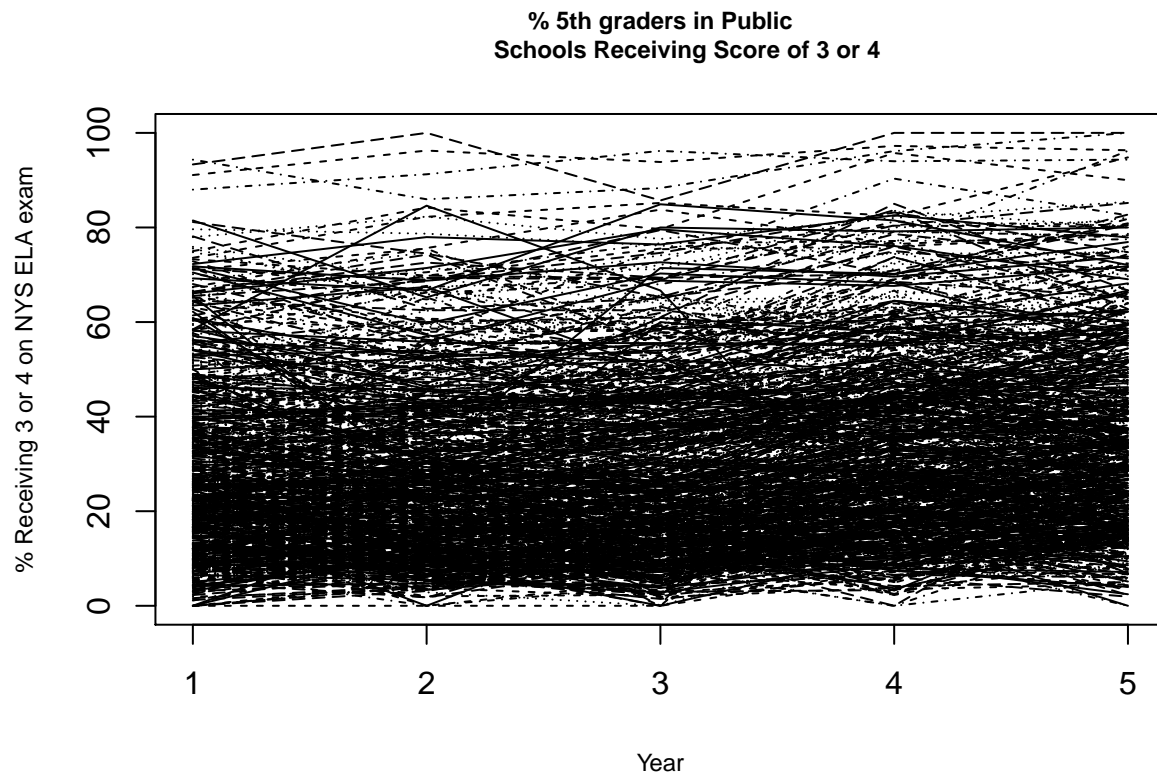

Then, I ran Mclust to determine the best option for number of clusters for my data.

```r
require(mclust)
```

```
## Loading required package: mclust
```

```
## Package 'mclust' version 5.4
## Type 'citation("mclust")' for citing this R package in publications.
```

```r
mcl <- Mclust(ela.na.5[,c(8,12,16,20,24)])
summary(mcl)
```

```
## ----------------------------------------------------
```

```
## Gaussian finite mixture model fitted by EM algorithm
## ----------------------------------------------------
##
## Mclust VEE (ellipsoidal, equal shape and orientation) model with 4 components:
##
##  log.likelihood   n df      BIC      ICL
##       -12737.52 696 41 -25743.39 -25996.38
##
## Clustering table:
##   1   2   3   4
## 342  75 172 107
```

MCL results: 4 clusters VEE best fits the data. BIC= -25743.39. The size of the clusters aren't too bad –
342, 75, 172, 107.

Now, a visualization of the data with the clusters labeled.

```r
matplot(t(ela.na.5[,c(8,12,16,20,24)]),type='l',col=(1:4)[mcl$class],xlab = "Year",
        ylab='% Receiving 3 or 4 on NYS ELA exam', main = "% 5th graders in Public Schools
        Receiving Score of 3 or 4", cex.lab = .75, cex.main = .75)
```
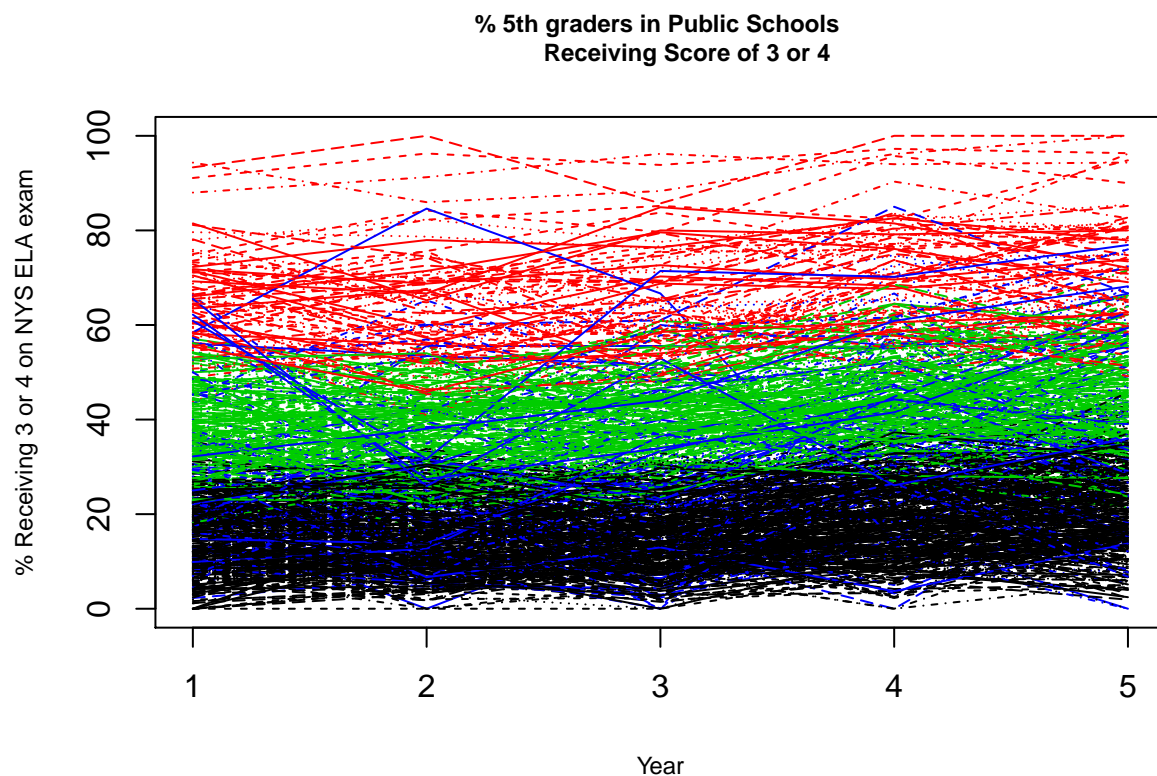

Model A1: Nagin clustering (Linear Term + Fixed)

Based on my results from mcl, I will run Nagin clustering with three clusters. I will include intercept, linear,
and quadratic terms to my model. To do this in R, I first need to reshape my dataset. I will also center my
Time variable.

```r
require(lcmm)
```

```
## Loading required package: lcmm
```

```
## Loading required package: survival
```

```r
ela.na.5.long <- reshape(ela.na.5[,c(1,8,12,16,20,24,25,26,27,28,29,30)],
                          varying=c("Three.Four.Per.2013","Three.Four.Per.2014","Three.Four.Per.2015",
                                    "Three.Four.Per.2016","Three.Four.Per.2017"),
                          v.names='Y',timevar="Time",idvar="DBN",direction="long")

#To get Time centered
for(i in 1:3480){
  ela.na.5.long$Time[i] <- ela.na.5.long$Time[i] - 3
}
```

One of the issues with the NYC DOE data is that each school is identified by a DBN code. However, these contain b oth characters and numbers, so I need an ID that contains only numbers for the R code to work properly. Once this is complete, I will order my data by ID and by Time. There are five total time points.

```r
uniqIDs <- unique(ela.na.5$DBN)
ela.na.5.long$ID <- match(ela.na.5.long$DBN,uniqIDs)
ela.na.5.long <- ela.na.5.long[order(ela.na.5.long$ID,ela.na.5.long$Time),]
```

Now that the data is in the appropriate format, I run the Nagin clustering.

```r
fit.nagin.fixed1 <- hlme(Y~Time,mixture=~Time,classmb=~1, subject='ID',ng=4,
                          data=ela.na.5.long)
```

```
## Be patient, hlme is running ...
## The program took 4.21 seconds
```

```r
summary(fit.nagin.fixed1)
```

```
## Heterogenous linear mixed model
##      fitted by maximum likelihood method
##
## hlme(fixed = Y ~ Time, mixture = ~Time, subject = "ID", classmb = ~1,
##      ng = 4, data = ela.na.5.long)
##
## Statistical Model:
##      Dataset: ela.na.5.long
##      Number of subjects: 696
##      Number of observations: 3480
##      Number of latent classes: 4
##      Number of parameters: 12
##
## Iteration process:
##      Convergence criteria satisfied
##      Number of iterations:  23
##      Convergence criteria: parameters= 1.7e-09
##                          : likelihood= 3.5e-10
##                          : second derivatives= 2.6e-15
##
## Goodness-of-fit statistics:
##      maximum log-likelihood: -13261.53
##      AIC: 26547.07
```

```
##      BIC: 26601.61
##
##
## Maximum Likelihood Estimates:
##
## Fixed effects in the class-membership model:
## (the class of reference is the last class)
##
##                    coef      Se    Wald p-value
## intercept class1  1.22137 0.15542   7.858 0.00000
## intercept class2  0.75364 0.14733   5.115 0.00000
## intercept class3  0.87301 0.14719   5.931 0.00000
##
## Fixed effects in the longitudinal model:
##
##                    coef      Se    Wald p-value
## intercept class1 14.32149 0.46982  30.483 0.00000
## intercept class2 44.12900 0.80825  54.598 0.00000
## intercept class3 27.62996 1.08383  25.493 0.00000
## intercept class4 67.38584 0.60369 111.624 0.00000
## Time class1       1.37538 0.18244   7.539 0.00000
## Time class2       2.46072 0.22558  10.908 0.00000
## Time class3       2.42513 0.22088  10.979 0.00000
## Time class4       2.14772 0.32104   6.690 0.00000
##
##                            coef      Se
## Residual standard error:  8.68912 0.11042
```

The equations for each cluster are:

Cluster 1: $14.32 + 1.38(Time)$

Cluster 2: $27.63 + 2.43(Time)$

Cluster 3: $44.13 + 2.46(Time)$

Cluster 4: $67.39 + 2.15(Time)$

BIC: 26601.61

Both the intercept and linear term are significant for all four clusters.

The residual standard error coefficient seems quite high: 8.69.

**NOTE: the green and red (clusters 2 and 3) clusters label switch a few times. All clusters will be reported in intercept-ascending order for clarity.

```r
#xlim is default
plot(fit.nagin.fixed1,which='fit',var.time='Time',legend=NULL, break.times=c(-2,-1,0,1,2),
    xlim=c(-1.5,1.5))
```

Weighted marginal predictions
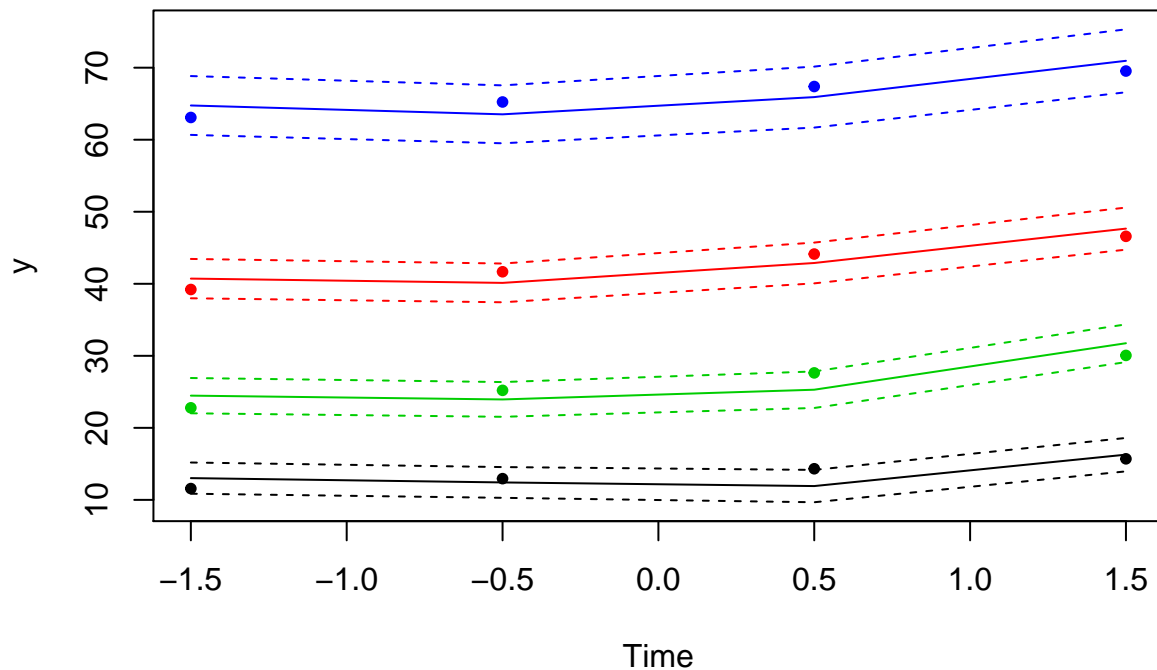
Model A2: Nagin clustering (Quadratic Term + Fixed)

```
fit.nagin.fixed2 <- hlme(Y~Time+I(Time^2),mixture=~Time+I(Time^2),classmb=~1,
                         subject='ID',ng=4, data=ela.na.5.long)
```

```
## Be patient, hlme is running ...
## The program took 4.97 seconds
```

```
summary(fit.nagin.fixed2)
```

```
## Heterogenous linear mixed model
##      fitted by maximum likelihood method
##
## hlme(fixed = Y ~ Time + I(Time^2), mixture = ~Time + I(Time^2),
##     subject = "ID", classmb = ~1, ng = 4, data = ela.na.5.long)
##
## Statistical Model:
##      Dataset: ela.na.5.long
##      Number of subjects: 696
##      Number of observations: 3480
##      Number of latent classes: 4
##      Number of parameters: 16
##
## Iteration process:
##      Convergence criteria satisfied
```

```
##        Number of iterations:  16
##        Convergence criteria: parameters= 2.8e-07
##                            : likelihood= 8.9e-08
##                            : second derivatives= 2.4e-13
##
## Goodness-of-fit statistics:
##        maximum log-likelihood: -13239.53
##        AIC: 26511.07
##        BIC: 26583.79
##
##
## Maximum Likelihood Estimates:
##
## Fixed effects in the class-membership model:
## (the class of reference is the last class)
##
##                      coef      Se   Wald p-value
## intercept class1  1.20612 0.15038  8.020 0.00000
## intercept class2  0.86901 0.14580  5.960 0.00000
## intercept class3  0.75786 0.14593  5.193 0.00000
##
## Fixed effects in the longitudinal model:
##
##                      coef      Se   Wald p-value
## intercept class1 12.89853 0.52097 24.758 0.00000
## intercept class2 26.28217 1.12459 23.370 0.00000
## intercept class3 43.10705 0.75782 56.883 0.00000
## intercept class4 66.35400 0.79098 83.888 0.00000
## Time class1        1.36798 0.18167  7.530 0.00000
## Time class2        2.42508 0.21924 11.061 0.00000
## Time class3        2.45720 0.22245 11.046 0.00000
## Time class4        2.14783 0.31723  6.771 0.00000
## I(Time^2) class1  0.67794 0.14586  4.648 0.00000
## I(Time^2) class2  0.58341 0.18373  3.175 0.00150
## I(Time^2) class3  0.44360 0.18558  2.390 0.01683
## I(Time^2) class4  0.48546 0.26258  1.849 0.06449
##
##                                 coef      Se
## Residual standard error:  8.62627 0.10914
```

The equations for each cluster are:

Cluster 1: $12.90 + 1.37(Time) + 0.68(Time)^2$

Cluster 2: $26.28 + 2.43(Time) + 0.58(Time)^2$

Cluster 3: $43.11 + 2.46(Time) + 0.44(Time)^2$

Cluster 4: $66.35 + 2.15(Time) + 0.49(Time)^2$

BIC: 26583.79

*Note that Cluster 2 and Cluster 3 labels switched. Corrected in the order of equations above.

Yet not all of the components in each equation are significant. For clusters 1, 2, and 3, all terms are significant. For cluster 4, all terms except the quadratic are significant (though this misses out just barely).

The BIC has decreased from the linear fixed model.

```
#Visualize the results
plot(fit.nagin.fixed2,which='fit',var.time='Time',legend=NULL, break.times=c(-2,-1,0,1,2),
     xlim=c(-1.5,1.5))
```

Weighted marginal predictions
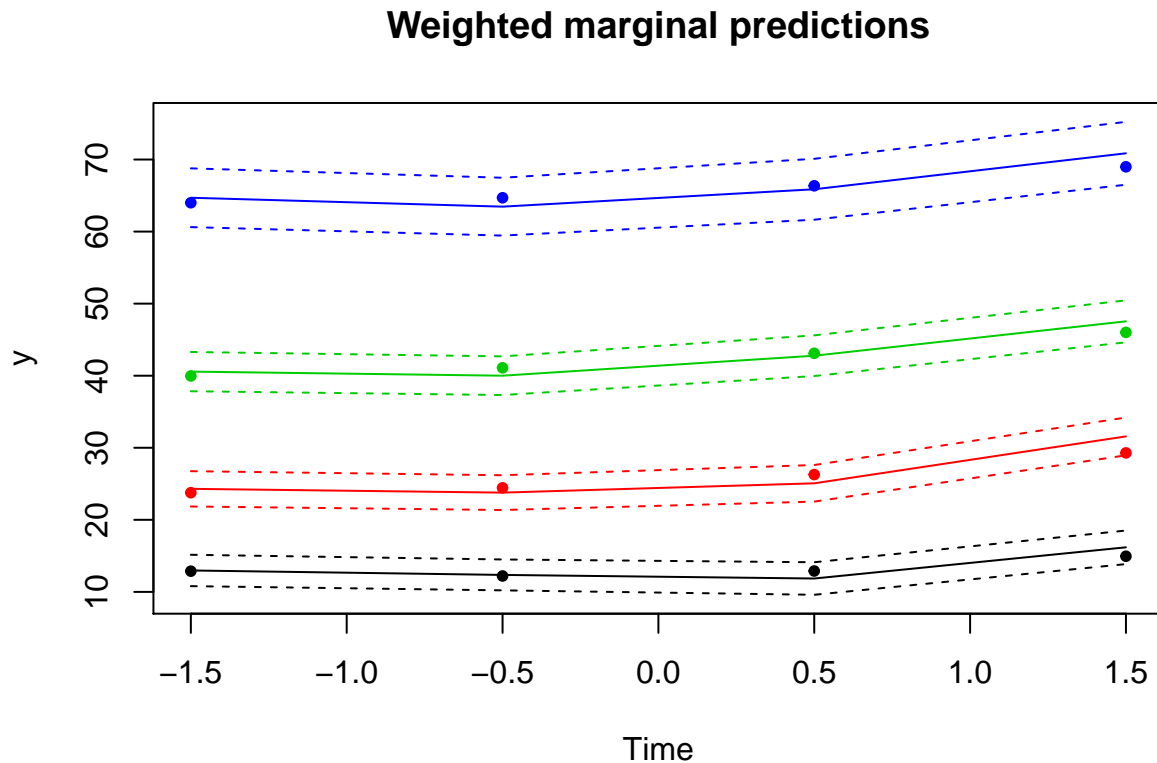
Model B1: Nagin clustering (Linear Term + Random Intercept)

The following three models will now use random intercept, which in the R code is the random argument.

```
fit.nagin.ri1 <- hlme(Y~Time,mixture=~Time,classmb=~1, random=~1,
                      subject='ID',ng=4, data=ela.na.5.long)
```

```
## Be patient, hlme is running ...
## The program took 3.91 seconds
```

```
summary(fit.nagin.ri1)
```

```
## Heterogenous linear mixed model
##      fitted by maximum likelihood method
##
## hlme(fixed = Y ~ Time, mixture = ~Time, random = ~1, subject = "ID",
##     classmb = ~1, ng = 4, data = ela.na.5.long)
##
## Statistical Model:
##      Dataset: ela.na.5.long
##      Number of subjects: 696
##      Number of observations: 3480
```

9

```
##      Number of latent classes: 4
##      Number of parameters: 13
##
## Iteration process:
##      Convergence criteria satisfied
##      Number of iterations:  18
##      Convergence criteria: parameters= 6.4e-05
##                          : likelihood= 8.7e-06
##                          : second derivatives= 1e-11
##
## Goodness-of-fit statistics:
##      maximum log-likelihood: -12973.73
##      AIC: 25973.45
##      BIC: 26032.54
##
##
## Maximum Likelihood Estimates:
##
## Fixed effects in the class-membership model:
## (the class of reference is the last class)
##
##                    coef      Se   Wald p-value
## intercept class1  2.01208 0.21883  9.195 0.00000
## intercept class2  1.25342 0.23154  5.413 0.00000
## intercept class3 -0.28971 0.37120 -0.780 0.43511
##
## Fixed effects in the longitudinal model:
##
##                    coef      Se   Wald p-value
## intercept class1 18.91941 0.57834 32.713 0.00000
## intercept class2 44.65717 1.28929 34.637 0.00000
## intercept class3 34.32082 2.40225 14.287 0.00000
## intercept class4 70.52143 2.03491 34.656 0.00000
## Time class1       1.46270 0.13305 10.994 0.00000
## Time class2       1.77639 0.25225  7.042 0.00000
## Time class3       7.63798 0.70056 10.903 0.00000
## Time class4       2.61859 0.41304  6.340 0.00000
##
##
## Variance-covariance matrix of the random-effects:
##          intercept
## intercept  59.76507
##
##                         coef      Se
## Residual standard error:  7.09081 0.10168
```

The equations for each cluster are:

Cluster 1: $18.92 + 1.46(Time)$

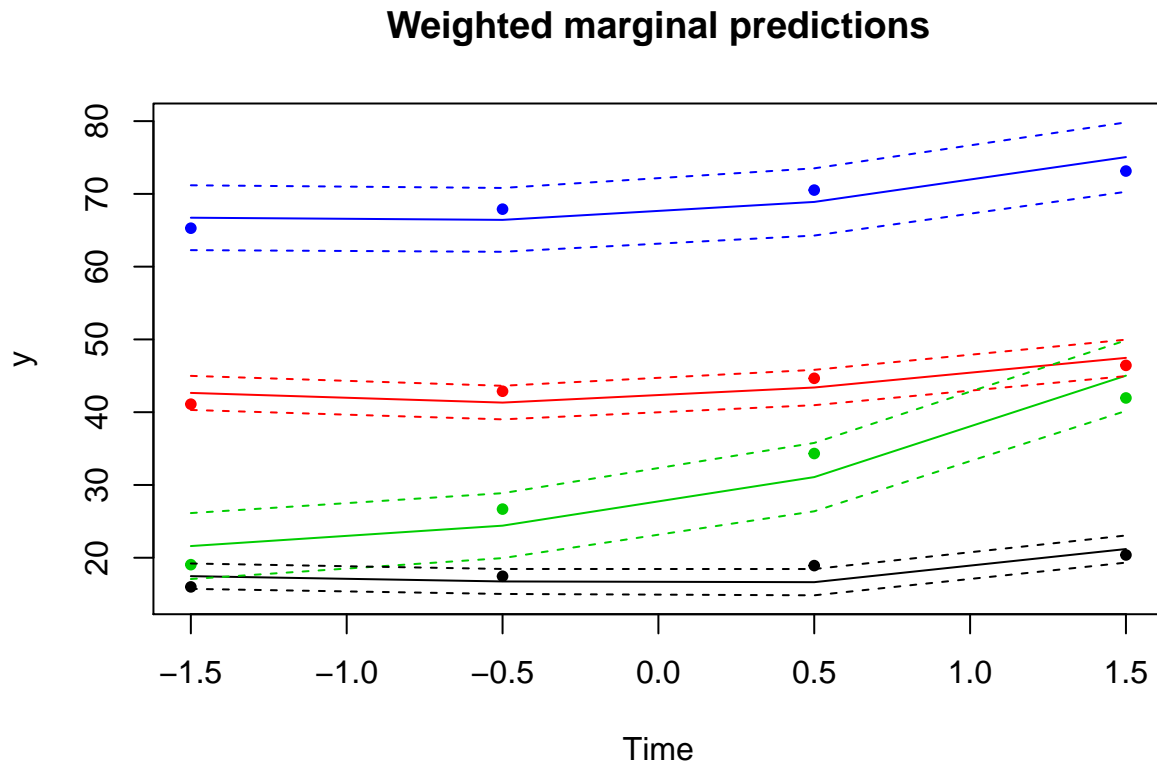Cluster 2: $34.32 + 7.64(Time)$

Cluster 3: $44.66 + 1.78(Time)$

Cluster 4: $70.52 + 2.62(Time)$

BIC: 26032.54

All terms for each cluster are significant.

```
plot(fit.nagin.ri1,which='fit',var.time='Time',legend=NULL, break.times=c(-2,-1,0,1,2),
     xlim=c(-1.5,1.5))
```

Weighted marginal predictions

The graph tells a much different story from the fixed effects clustering. There is clearly some curvature, suggesting the need for higher-order terms. Our four cluster trajectories are also no longer "separate." Rather, the green trajectory (cluster 2) is moves closer over time to the red trajectory (cluster 3).

Model B2: Nagin clustering (Quadratic Term + Random Intercept)

```
fit.nagin.ri2 <- hlme(Y~Time+I(Time^2),mixture=~Time+I(Time^2),classmb=~1, random=~1,
                      subject='ID',ng=4, data=ela.na.5.long)
```

```
## Be patient, hlme is running ...
## The program took 10.21 seconds
```

```
summary(fit.nagin.ri2)
```

```
## Heterogenous linear mixed model
##      fitted by maximum likelihood method
##
## hlme(fixed = Y ~ Time + I(Time^2), mixture = ~Time + I(Time^2),
##     random = ~1, subject = "ID", classmb = ~1, ng = 4, data = ela.na.5.long)
##
## Statistical Model:
```

11

```
##          Dataset: ela.na.5.long
##          Number of subjects: 696
##          Number of observations: 3480
##          Number of latent classes: 4
##          Number of parameters: 17
##
## Iteration process:
##          Convergence criteria satisfied
##          Number of iterations:  30
##          Convergence criteria: parameters= 2.8e-08
##                              : likelihood= 4.4e-09
##                              : second derivatives= 3.7e-14
##
## Goodness-of-fit statistics:
##          maximum log-likelihood: -12939.94
##          AIC: 25913.88
##          BIC: 25991.15
##
##
## Maximum Likelihood Estimates:
##
## Fixed effects in the class-membership model:
## (the class of reference is the last class)
##
##                          coef       Se    Wald p-value
## intercept class1   2.00910 0.22114   9.085 0.00000
## intercept class2  -0.27056 0.38348  -0.706 0.48047
## intercept class3   1.26423 0.23384   5.406 0.00000
##
## Fixed effects in the longitudinal model:
##
##                          coef       Se    Wald p-value
## intercept class1 17.62463 0.62605 28.152 0.00000
## intercept class2 31.62141 2.79182 11.326 0.00000
## intercept class3 43.78408 1.29977 33.686 0.00000
## intercept class4 69.83177 2.23295 31.273 0.00000
## Time class1        1.44168 0.13315 10.828 0.00000
## Time class2        7.62517 0.71049 10.732 0.00000
## Time class3        1.79560 0.26549  6.763 0.00000
## Time class4        2.63209 0.41661  6.318 0.00000
## I(Time^2) class1   0.63347 0.09829  6.445 0.00000
## I(Time^2) class2   0.99714 0.44337  2.249 0.02451
## I(Time^2) class3   0.42080 0.15832  2.658 0.00786
## I(Time^2) class4   0.35425 0.28528  1.242 0.21433
##
##
## Variance-covariance matrix of the random-effects:
##            intercept
## intercept  60.31125
##
##                                 coef       Se
## Residual standard error:  6.99699 0.10104
```

The equations for each cluster are:

Cluster 1: $17.62 + 1.44(Time) + 0.63(Time)^2$

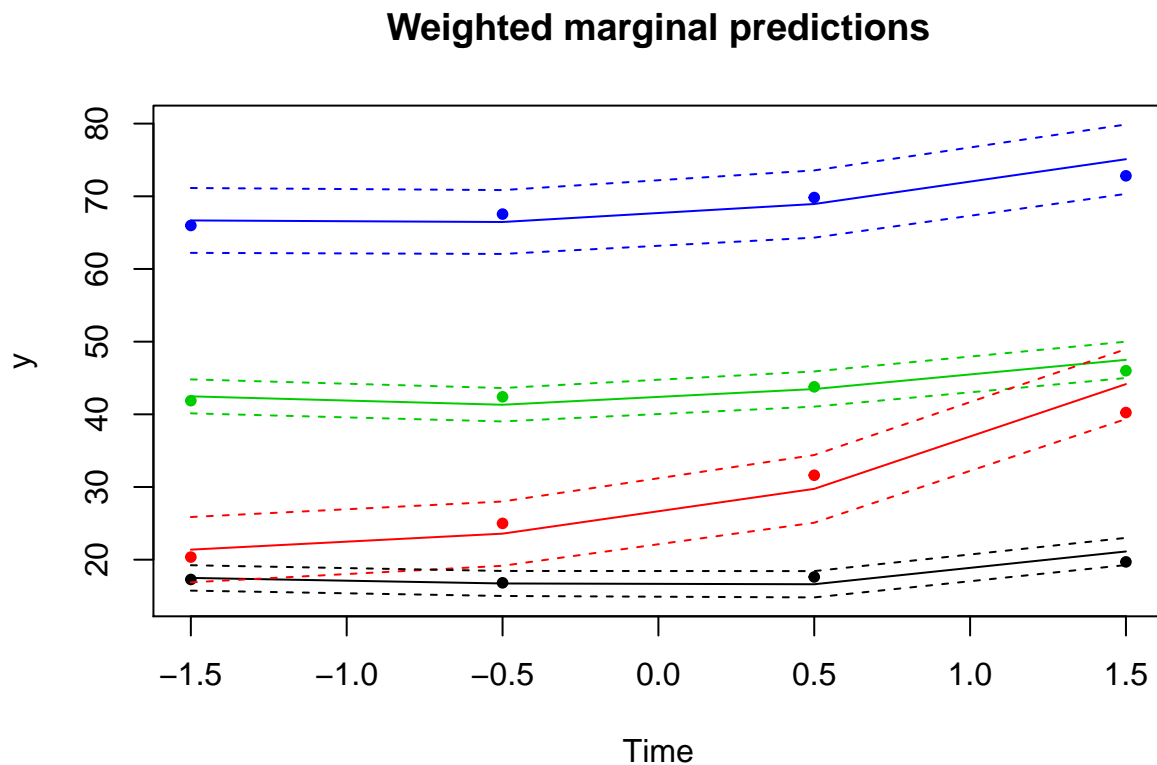Cluster 2: $31.62 + 7.63(Time) + 1.0(Time)^2$

Cluster 3: $43.79 + 1.80(Time) + 0.42(Time)^2$

Cluster 4: $69.83 + 2.63(Time) + 0.35(Time)^2$

BIC: 25991.15

For clusters 1, 2, and 3, all three terms are significant. The quadratic term is not significant for cluster 4.

```
plot(fit.nagin.ri2,which='fit',var.time='Time',legend=NULL, break.times=c(-2,-1,0,1,2),
     xlim=c(-1.5,1.5))
```

Weighted marginal predictions

Model C1: Nagin clustering (Linear Term + Random Slope)

```
fit.nagin.rs1 <- hlme(Y~Time,mixture=~Time,classmb=~1, random=~Time, subject='ID',ng=4,
                      data=ela.na.5.long)
```

```
## Be patient, hlme is running ...
## The program took 5.35 seconds
```

```
summary(fit.nagin.rs1)
```

```
## Heterogenous linear mixed model
##      fitted by maximum likelihood method
##
```

```
## hlme(fixed = Y ~ Time, mixture = ~Time, random = ~Time, subject = "ID",
##     classmb = ~1, ng = 4, data = ela.na.5.long)
##
## Statistical Model:
##     Dataset: ela.na.5.long
##     Number of subjects: 696
##     Number of observations: 3480
##     Number of latent classes: 4
##     Number of parameters: 15
##
## Iteration process:
##     Convergence criteria satisfied
##     Number of iterations:  19
##     Convergence criteria: parameters= 1.8e-06
##                         : likelihood= 1.1e-06
##                         : second derivatives= 9.8e-14
##
## Goodness-of-fit statistics:
##     maximum log-likelihood: -12955.83
##     AIC: 25941.67
##     BIC: 26009.85
##
##
## Maximum Likelihood Estimates:
##
## Fixed effects in the class-membership model:
## (the class of reference is the last class)
##
##                       coef      Se   Wald p-value
## intercept class1   1.96707 0.17896 10.992 0.00000
## intercept class2  -1.48153 0.51879 -2.856 0.00429
## intercept class3   1.24906 0.18939  6.595 0.00000
##
## Fixed effects in the longitudinal model:
##
##                       coef      Se   Wald p-value
## intercept class1 19.06226 0.55533 34.326 0.00000
## intercept class2 36.21732 3.65092  9.920 0.00000
## intercept class3 43.94263 1.00077 43.909 0.00000
## intercept class4 69.55752 1.53718 45.250 0.00000
## Time class1       1.81226 0.15050 12.042 0.00000
## Time class2       9.81627 1.29838  7.560 0.00000
## Time class3       2.07763 0.21813  9.525 0.00000
## Time class4       1.32940 0.42580  3.122 0.00180
##
##
## Variance-covariance matrix of the random-effects:
##           intercept    Time
## intercept  59.81108
## Time        6.48133 2.02041
##
##                             coef      Se
## Residual standard error:  6.88159 0.10651
```

The equations for each cluster are:

Cluster 1: $19.06 + 1.81(Time)$
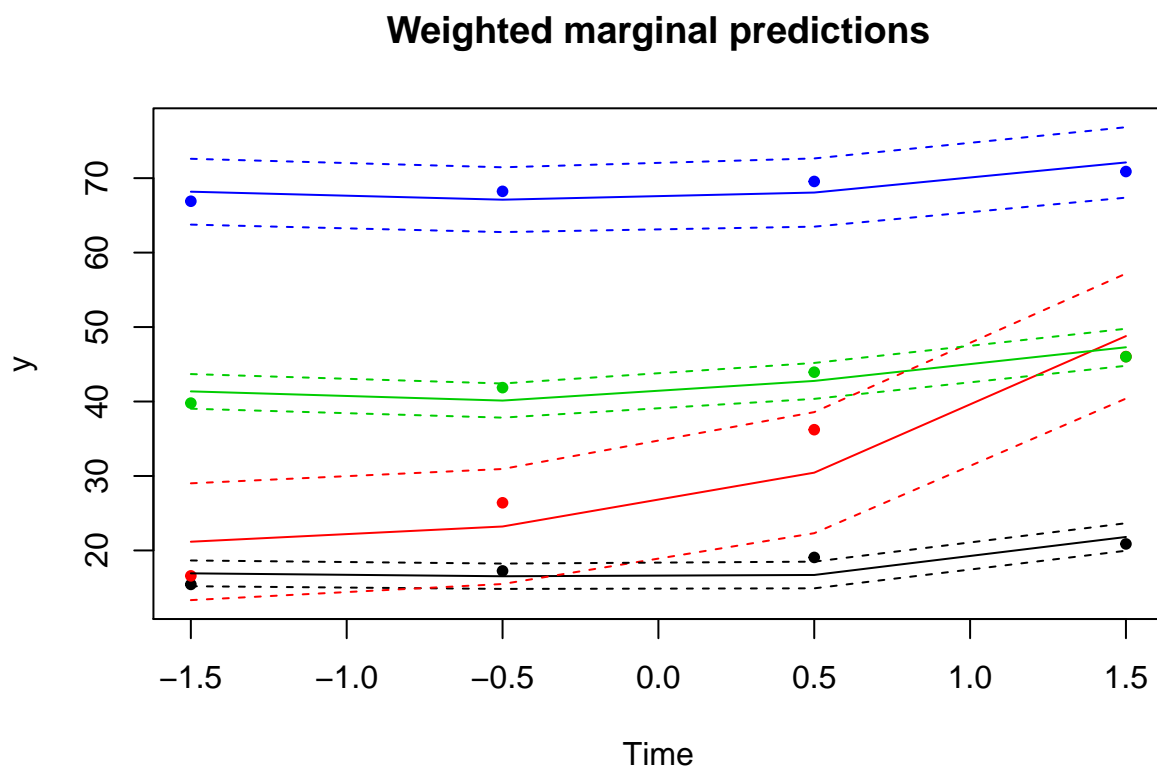
Cluster 2: $36.22 + 9.82(Time)$

Cluster 3: $43.94 + 2.10(Time)$

Cluster 4: $69.56 + 1.33(Time)$

BIC: 26009.85

All terms are significant for each cluster. The residual standard error coefficient is 6.88, which is much lower than that of the fixed linear.

```
plot(fit.nagin.rs1,which='fit',var.time='Time',legend=NULL, break.times=c(-2,-1,0,1,2),
     xlim=c(-1.5,1.5))
```

Weighted marginal predictions

Model C2: Nagin clustering (Quadratic Term + Random Slope)

```
fit.nagin.rs2 <- hlme(Y~Time+I(Time^2),mixture=~Time+I(Time^2),classmb=~1, random=~Time, subject='ID',
                 ng=4, data=ela.na.5.long)
```

```
## Be patient, hlme is running ...
## The program took 8.13 seconds
```

```
summary(fit.nagin.rs2)
```

```
## Heterogenous linear mixed model
##       fitted by maximum likelihood method
##
## hlme(fixed = Y ~ Time + I(Time^2), mixture = ~Time + I(Time^2),
##      random = ~Time, subject = "ID", classmb = ~1, ng = 4, data = ela.na.5.long)
##
## Statistical Model:
##       Dataset: ela.na.5.long
##       Number of subjects: 696
##       Number of observations: 3480
##       Number of latent classes: 4
##       Number of parameters: 19
##
## Iteration process:
##       Convergence criteria satisfied
##       Number of iterations:  18
##       Convergence criteria: parameters= 2.3e-06
##                           : likelihood= 1.1e-06
##                           : second derivatives= 6.6e-14
##
## Goodness-of-fit statistics:
##       maximum log-likelihood: -12913.79
##       AIC: 25865.58
##       BIC: 25951.95
##
##
## Maximum Likelihood Estimates:
##
## Fixed effects in the class-membership model:
## (the class of reference is the last class)
##
##                    coef      Se    Wald p-value
## intercept class1  1.97013 0.17767 11.089 0.00000
## intercept class2  1.26171 0.18748  6.730 0.00000
## intercept class3 -1.79787 0.46948 -3.830 0.00013
##
## Fixed effects in the longitudinal model:
##
##                     coef      Se    Wald p-value
## intercept class1 17.86645 0.59229 30.165 0.00000
## intercept class2 43.10763 1.01063 42.654 0.00000
## intercept class3 28.70555 4.36723  6.573 0.00000
## intercept class4 68.70200 1.64959 41.648 0.00000
## Time class1       1.84704 0.14538 12.705 0.00000
## Time class2       2.11861 0.21606  9.805 0.00000
## Time class3      10.67693 1.21252  8.806 0.00000
## Time class4       1.27520 0.42626  2.992 0.00278
## I(Time^2) class1  0.62893 0.09325  6.745 0.00000
## I(Time^2) class2  0.39264 0.14109  2.783 0.00539
## I(Time^2) class3  3.10454 0.72325  4.292 0.00002
## I(Time^2) class4  0.40522 0.25420  1.594 0.11090
##
##
## Variance-covariance matrix of the random-effects:
```

```
##           intercept     Time
## intercept  61.30367
## Time         7.05425 2.34054
##
##                          coef      Se
## Residual standard error:  6.73310 0.10460
```

The equations for each cluster are:

Cluster 1: $17.87 + 1.85(Time) + 0.63(Time)^2$

Cluster 2: $28.71 + 10.68(Time) + 0.39(Time)^2$

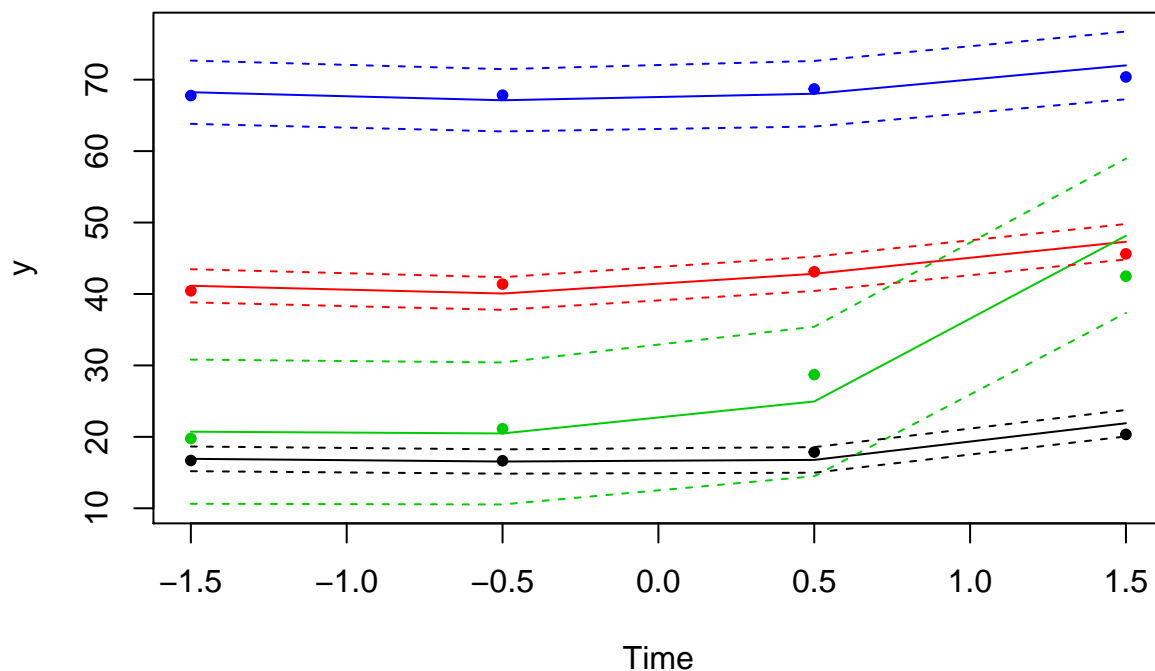Cluster 3: $43.11 + 2.12(Time) + 0.31(Time)^2$

Cluster 4: $68.70 + 1.28(Time) + 0.41(Time)^2$

BIC: 25951.95

For clusters 1, 2, and 3, all three terms are significant. For cluster 4, the quadratic term is not significant.

```
plot(fit.nagin.rs2,which='fit',var.time='Time',legend=NULL, break.times=c(-2,-1,0,1,2),
     xlim=c(-1.5,1.5))
```

Weighted marginal predictions

Results: Which model(s) is best?

Since the hlme package does not allow for different order terms, I must take into account that while the higher-order term models have lower BICs, this may be because only some of the clusters require higher-order

terms, but not others.

The BICs for the 6 models are:

A1: 26601.61

A2: 26583.79

B1: 26032.54

B2: 25991.15

C1: 26009.85

C2: 25951.95

The simplest model (A1) has the highest BIC. However, since the quadratic terms were significant for 3 out of the 4 clusters, thus indicating a quadratic means structure, it is appropriate to choose A2 as the best model.

Model D1: A2 + Risk Factor(Brooklyn)

Since A2 is one of the better model choices, I will add a time-invariant risk factor (Brooklyn; Not Brooklyn) to my model to see if BIC improves.

```
fit.nagin.cov4 <- hlme(Y~Time+I(Time^2),mixture=~Time+I(Time^2),classmb=~ela.na.5.long$Brooklyn,
                       subject='ID',ng=4, data=ela.na.5.long)
```

```
## Be patient, hlme is running ...
## The program took 10.06 seconds
```

```
summary(fit.nagin.cov4)
```

```
## Heterogenous linear mixed model
##      fitted by maximum likelihood method
##
## hlme(fixed = Y ~ Time + I(Time^2), mixture = ~Time + I(Time^2),
##      subject = "ID", classmb = ~ela.na.5.long$Brooklyn, ng = 4,
##      data = ela.na.5.long)
##
## Statistical Model:
##      Dataset: ela.na.5.long
##      Number of subjects: 696
##      Number of observations: 3480
##      Number of latent classes: 4
##      Number of parameters: 19
##
## Iteration process:
##      Convergence criteria satisfied
##      Number of iterations:  24
##      Convergence criteria: parameters= 7.6e-08
##                          : likelihood= 8.5e-08
##                          : second derivatives= 1.1e-14
##
## Goodness-of-fit statistics:
##      maximum log-likelihood: -13203.51
##      AIC: 26445.02
##      BIC: 26531.38
##
```

```
##
## Maximum Likelihood Estimates:
##
## Fixed effects in the class-membership model:
## (the class of reference is the last class)
##
##                                    coef      Se    Wald p-value
## intercept class1               2.01921 0.19948 10.122 0.00000
## intercept class2               1.41828 0.21026  6.745 0.00000
## intercept class3               1.09653 0.21819  5.026 0.00000
## ela.na.5.long$Brooklyn1 class1  0.59295 0.41913  1.415 0.15715
## ela.na.5.long$Brooklyn1 class2  0.75366 0.43176  1.746 0.08089
## ela.na.5.long$Brooklyn1 class3  0.33834 0.45560  0.743 0.45771
##
## Fixed effects in the longitudinal model:
##
##                      coef      Se    Wald p-value
## intercept class1 14.81656 0.38478 38.507 0.00000
## intercept class2 33.41236 0.66332 50.372 0.00000
## intercept class3 50.96864 0.69031 73.834 0.00000
## intercept class4 74.81153 1.05560 70.871 0.00000
## Time class1       1.53894 0.15157 10.153 0.00000
## Time class2       2.46618 0.20464 12.052 0.00000
## Time class3       2.42223 0.25580  9.469 0.00000
## Time class4       2.36457 0.45639  5.181 0.00000
## I(Time^2) class1  0.69152 0.12771  5.415 0.00000
## I(Time^2) class2  0.39459 0.17158  2.300 0.02146
## I(Time^2) class3  0.64016 0.21282  3.008 0.00263
## I(Time^2) class4  0.24540 0.37713  0.651 0.51523
##
##                      coef      Se
## Residual standard error:  8.67236 0.10669
```

```r
plot(fit.nagin.cov4,which='fit',var.time='Time',legend=NULL, break.times=c(-2,-1,0,1,2),
     xlim=c(-1.5,1.5))
```

Weighted marginal predictions
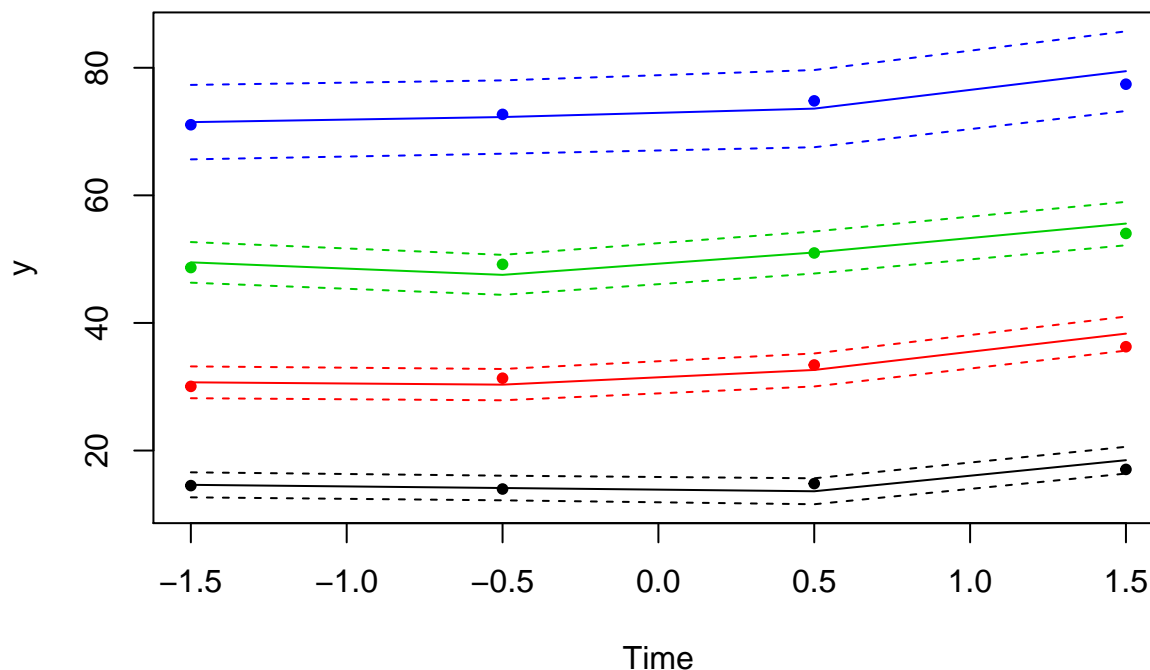
The equations for each cluster:

Cluster 1: $14.82 + 1.54(Time) + 0.69(Time)^2$

Cluster 2: $33.41 + 2.50(Time) + 0.39(Time)^2$

Cluster 3: $50.97 + 2.42(Time) + 0.64(Time)^2$

Cluster 4: $74.81 + 2.36(Time) + 0.25(Time)^2$

BIC: 26531.38

The BIC is pretty close to that of A2. Although the BIC is about 70 points lower than A2's BIC, D1 is a more complex model. From the output, we can see that roughly, not being located in Brooklyn is associated with a greater likelihood of being placed in Clusters 1 (7.53 times higher), 2 (4.13 times higher), or 3 (2.99 times higher), all compared to Cluster 4, the highest performing group. However, the coefficients for Brooklyn are not statistically significant.

These results should be taken with caution due to the fact that not all of the city's public schools were represented in this dataset. Additionally, Brooklyn has an ever-increasing number of charter schools, often paired with public school closures. Thus, the above model may indicate different results when used on the full NYC public elementary school population.

Therefore, I consider this model (D1) not as the best fit for the data, but as a strong secondary option.

K-means cluster solution

For quick comparison, I used NbClust to see how many clusters would yield the best C(g) for k-means, a non model-based method.

```r
library(NbClust)
k.means <- NbClust(ela.na.5[,c(8,12,16,20,24)],method='kmeans',index='ch')
k.means$Best.nc
```

```
## Number_clusters      Value_Index
##           2.000         1134.938
```

```r
#k-means with 100 random starts
set.seed(2011)
km.2 <- kmeans(ela.na.5[,c(8,12,16,20,24)], 2, nstart= 100)
set.seed(2011)
km.3 <- kmeans(ela.na.5[,c(8,12,16,20,24)], 3, nstart= 100)
set.seed(2011)
km.4 <- kmeans(ela.na.5[,c(8,12,16,20,24)], 4, nstart= 100)
set.seed(2011)
km.5 <- kmeans(ela.na.5[,c(8,12,16,20,24)], 5, nstart= 100)


c2 <- c.crit(km.2)
c3 <- c.crit(km.3)
c4 <- c.crit(km.4)
c5 <- c.crit(km.5)
c2$C.g
```

```
## [1] 1134.939
```

```r
c3$C.g
```

```
## [1] 1103.878
```

```r
c4$C.g
```

```
## [1] 990.4247
```

```r
c5$C.g
```

```
## [1] 907.2167
```

Suprisingly, k-means suggests a two cluster solution. This is important to keep in mind. I'll check what the hlme looks like with only two clusters to see if it mirrors the k-means results.

```r
fixed.nagin.two <- hlme(Y~Time,mixture=~Time,classmb=~1, subject='ID',ng=2, data=ela.na.5.long)
```

```
## Be patient, hlme is running ...
## The program took 0.69 seconds
```

```r
summary(fixed.nagin.two)
```

```
## Heterogenous linear mixed model
##      fitted by maximum likelihood method
##
## hlme(fixed = Y ~ Time, mixture = ~Time, subject = "ID", classmb = ~1,
##     ng = 2, data = ela.na.5.long)
##
## Statistical Model:
##      Dataset: ela.na.5.long
##      Number of subjects: 696
##      Number of observations: 3480
##      Number of latent classes: 2
##      Number of parameters: 6
```

```
##
## Iteration process:
##      Convergence criteria satisfied
##      Number of iterations:  12
##      Convergence criteria: parameters= 4e-05
##                          : likelihood= 6.8e-05
##                          : second derivatives= 4.6e-10
##
## Goodness-of-fit statistics:
##      maximum log-likelihood: -14011.82
##      AIC: 28035.64
##      BIC: 28062.91
##
##
## Maximum Likelihood Estimates:
##
## Fixed effects in the class-membership model:
## (the class of reference is the last class)
##
##                      coef      Se   Wald p-value
## intercept class1  0.68084 0.08922  7.631 0.00000
##
## Fixed effects in the longitudinal model:
##
##                      coef      Se   Wald p-value
## intercept class1 20.19451 0.33277 60.686 0.00000
## intercept class2 52.22735 0.53241 98.096 0.00000
## Time class1       1.81323 0.17878 10.142 0.00000
## Time class2       2.37760 0.25280  9.405 0.00000
##
##                                coef      Se
## Residual standard error: 12.06326 0.14677
```
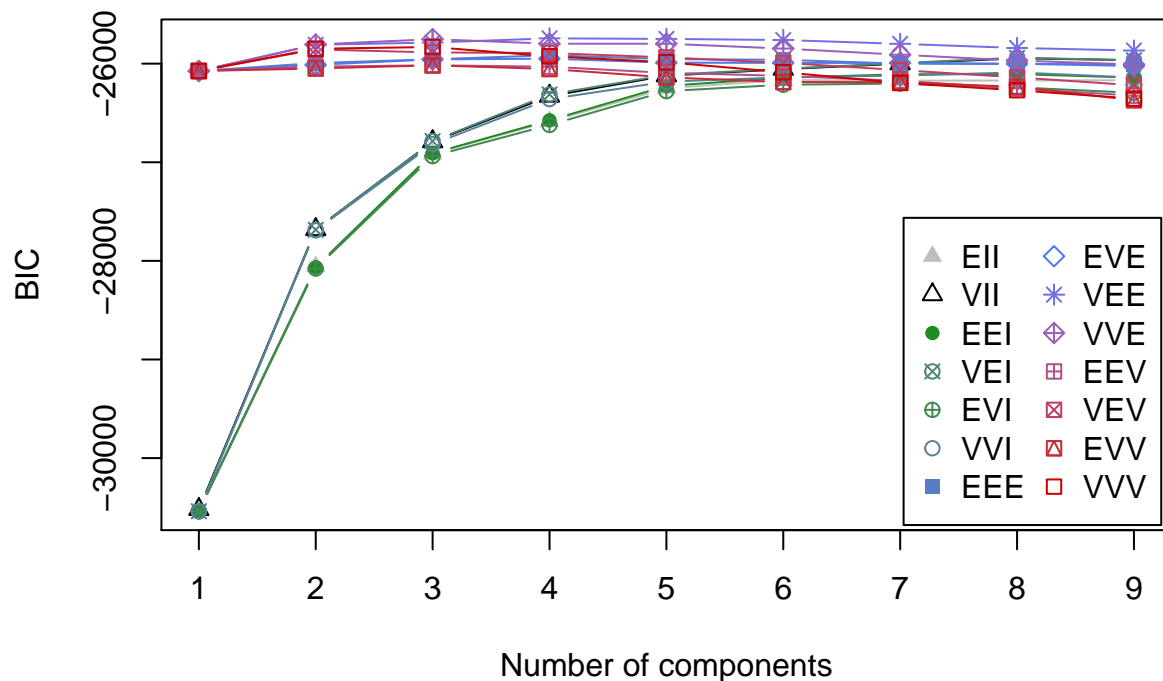
The BIC for this model is unexpectedly higher than any of the other models. Yet Mclust reported that the best solution is a 4 cluster solution. When further examining the BIC plot provided by Mclust, it is clear that there is some close competition for the best model around 2, 3, and 4 cluster solutions.

```
plot(mcl$BIC)
```

It is possible that I am getting different results because of how the data is restructured for hlme/nagin clustering. The five separate score variables are funneled into one "Time" variable, which is centered in the ELA long dataset. However, there is no "Time" variable in the original ELA dataset used for k-means. But this shouldn't cause too big a problem because the data are still the same. Nevertheless, I will re-run k-means with the elongated dataset.

```r
k1 <- kmeans(ela.na.5.long[,9], 4, nstart= 100)
k2 <- kmeans(ela.na.5.long[,9], 2, nstart= 100)
c.new1 <- c.crit(k1)
c.new2 <- c.crit(k2)
c.new1$C.g
```

```
## [1] 11722.03
```

```r
c.new2$C.g
```

```
## [1] 7837.113
```

Alas, the C(g) for the 4 cluster solution is greater than the C(g) for the two cluster solution, which is reassuring given the results from Mclust and Nagin clustering. Due to the structure of my data, I think it is best to stick with the previous Nagin clustering solutions, specifically model A2, than a k-means solution.