Detecting Learning Discontinuity for Out-of-tutor Events

Yiwen Zhang, Naifei Pan, Jie Luo

May 2021

Abstract

In this study, we addressed the question of how effective tutors are on students' performances in an online intelligent tutoring system. We used the data from datashop which recorded the 195 students' learning progress in an online math tutor program. We utilized the AFM model - a logistics regression- in our analyses to examine the effects of tutors' help. Our results showed that tutors' interventions had improved the students' performance in terms of the error rate. Also, students who received tutoring earlier also performed better than students who received tutoring later. This analysis could help improve the scientific understanding of learning with intelligent tutoring systems.

1 Introduction

In recent years, educational institutions have incorporated new technologies with traditional education to improve the overall learning experiences. With access to the internet and feasible devices, students can have a quality education wherever and whenever they want. On one hand, online education makes it easier to track students' progress as it records their performances in each pre-designed problem with relative knowledge. On the other hand, teachers are able to monitor the class through the screen and decide if additional help is needed for certain students. Thus, the effectiveness of educators' help can be reflected through the student's performances who receive the extra help. In this study, we seek to find out how educators' interventions affect students' learning progress on an online math tutoring system. Specifically, we will address the following questions:

1. Do these interventions put students on a different learning trajectory, with respect to the specific skills?

2. How can we measure the effect of teacher interventions on learning?

2 Data

The Out-of-tutor event detection data is provided by Datashop[**2**], which consists of 195 students' learning records on an online math tutoring system. There

are 3 sub-datasets, organized by transaction, student step, and student-problem. In this study, we used the Transaction data and Student Step data.

Each observation in the Transaction dataset is ordered by student and the transaction time, and the following shows the important variables that were measured.

Table 1. Description of important variables in Transaction dataset

Variable	Description
Row	A Row Counter
Anon Student Id	DataShop-generated anonymous student Id
Transaction Id	A Unique ID that identifies the transaction
Tutor Response Type	The type of response made by the tutor
Problem Name	The name of the problem
KC	The knowledge component for this transaction

Observations in Student Step dataset are ordered by student time of the first correct attempt (encoded as "Correct Transaction Time"). Detailed information for the variables is listed in table 2.

Table 2. Description of important variables in Student step dataset

Variable	Description
Anon Student Id	DataShop-generated anonymous student Id
First Attempt	The tutor's response to the student;s first attempt on the step. Example values are "hint","correct", and "incorrect"
Corrects	Total correct attempts by the student for the step
Problem Name	The name of the problem
KC	The knowledge component for this transaction
Opportunity	The first chance on a step for students to demonstrate whether they have learned the associated KC. Each time a student encounters a problem that has a listed KC, the opportunity number will increase by one

We map the tutor intervention as an indicator variable from the transaction dataset to the Student Step dataset, which indicates whether a tutor has intervened in the learning process during this observation. In this report, we choose two knowledge components for the analysis.

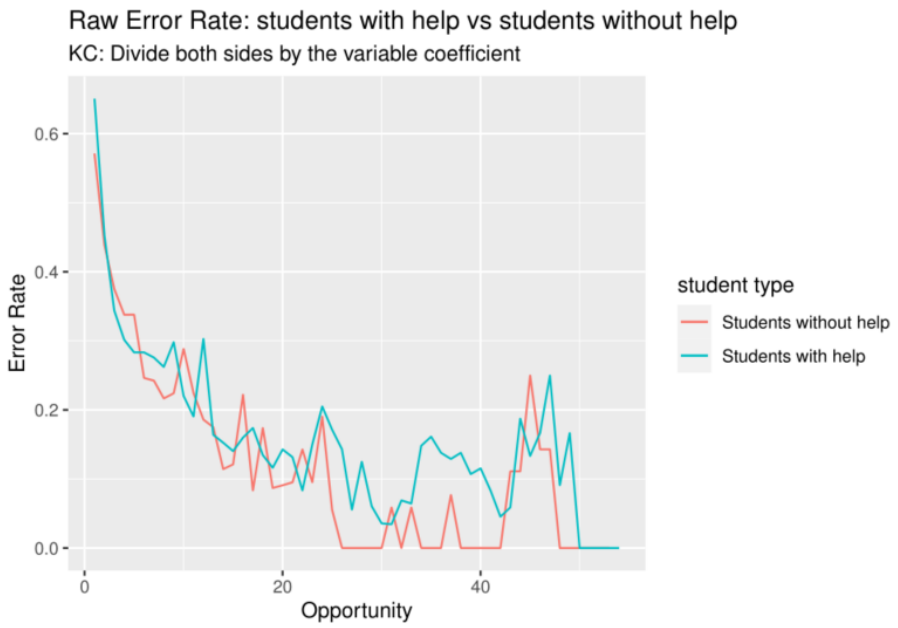


Figure 1: Raw error rate for Divide Both Sides By the Variable Coefficient

In Figure 1, we show the raw error rate for KC Divide Both Sides by The Variable Coefficient. There are 97 students who receive the help and 84 students who do not. From the plot, we do not observe an obvious difference on the raw error rate between students who receive tutor helps and students who do not receive tutor help in this KC.

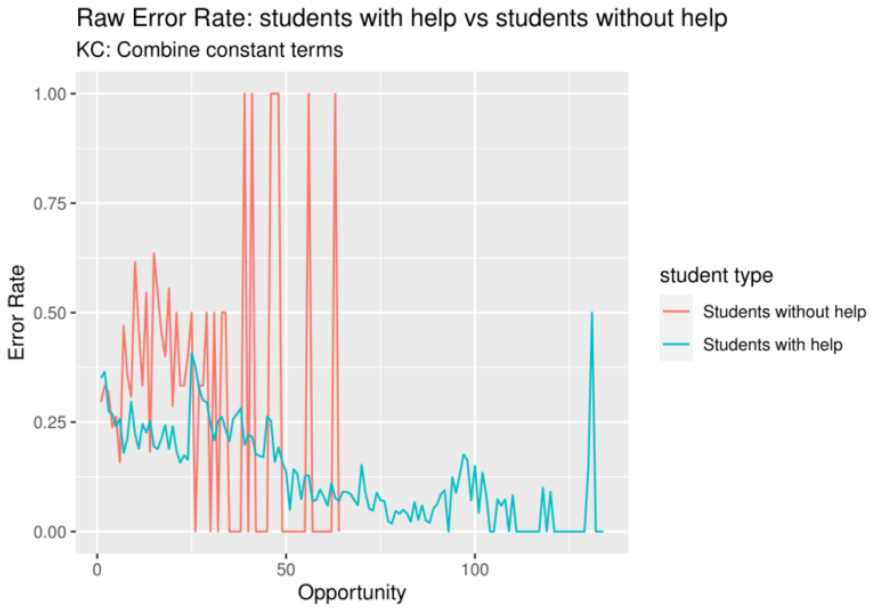


Figure 2: Raw error rate for Combine Constant Terms

Figure 2 shows the raw error rate for kc Combine Constant Terms. We notice that before opportunity 25, students without help have higher error rate than students without help. More details from exploratory data analysis can be found in Appendix 1.

3 Methods

3.1 Method 1

Our analysis has two parts. First, we make an adjustment on the original AFM model(Aleven Koedinger, 2013) and fit the new AFM model for each KC. For the original AFM model, we have:

$$ln\frac{p_{ij}}{1-p_{ij}} = \theta_i + \Sigma_k \beta_k Q_{kj} + \Sigma_k Q_{kj}(\gamma_k N_{ik})$$

In this equation, p represents the probability that student i gets the step j correct. θ represents students' initial proficiency. β represents the ease of KC k. γ represents gain for each opportunity to practice KC k. N represents the number of opportunities each student has to practice KC k, prior to step j. So, $\gamma * N$ will give us how much the student learned on prior opportunities for this KC. Q is an indicator variable which represents whether KC k is needed for step j.

In our new model, We assume one intervention only influences one KC. For each KC, we add the teacher intervention effect to the model.

$$ln\frac{p_{ij}}{1-p_{ij}} = \theta_i + \gamma_k N_{ik} + \phi_k N_{ik} I_{ik} + \alpha_k I_{ik}$$

The new AFM model also predicts the log likelihood that student i gets the correct answer. θ represents students' initial proficiency for certain KC, $\gamma * N$ represents how much the student learned on prior opportunities for this KC. I represents teacher intervention: whether the step is before or after first teacher intervention. Coefficient ϕ adjusts the learning rate based on the teacher intervention term, and the coefficient α adjusts the intercept based on the teacher intervention term. Therefore, if we observe a positive coefficient, that means teacher intervention may accelerate a student's learning rate, and vice versa.

To have a clear idea for pre-tutor and post-tutor performance, we also fit two separate original AFM using pre-tutor observations and post-tutor observations. More details from exploratory data analysis can be found in Appendix 2-5.

Below is a brief demonstration of how we define pre and post intervention opportunities.

Student 1	1	2	3	4	5
Student 2	1	2	3	4	5
Student 3	1	2	3	4	5

Figure 3: Illustration for data separation

For each KC, we check when the first intervention happens for each student. For example, in the table above, for student 1, the first intervention happens after the first opportunity. Therefore, the first opportunity is pre-tutor data and whatever after it is post-tutor data. For student 2, the first intervention happened after the third opportunity, so the first to the third opportunities are pre-tutor data, and whatever after the third opportunity are post-tutor data. The same logic applies to each student.

3.2 Method 2

Second, to validate our assumption about the results from method 1, we examine whether students who received teacher interventions at different times exhibit different learning rates. For each KC, we split the students into three groups based on the density plot of the intervention, which are the early-intervention group, normal-intervention group, and late-intervention group. We compare the raw error rate of these three groups, and the predicted error rate using a Group AFM model.

$$ln\frac{p_{ik}}{1-p_{ik}} = \theta_{ik} + \gamma_k N_{ik} + \psi_{km} N_{ik} G_{ik} + \gamma_{km} G_{ik}$$

Similar to our new AFM model. We use an indicator variable G, which represents the group the student belongs to: "Early", "Normal" or "Late". This model will help us know whether receiving a teacher early or late will have an effect on a student's learning rate.

4 Results

We implement the New AFM model using R. In this section we only present the results for two KCs.

4.1 Results from Method 1

4.1.1 KC: "Divide both sides by the variable coefficient"

The first KC is "Divide both sides by the variable coefficient". We separate all observations related to this KC into the Pre-tutor subset and Post-tutor subset, according to Method 1. We create the teacher indicator variable with "0" means pre-tutor and "1" represents post-tutor. After fitting the new AFM model, we compute the predicted error rate. To compare the predicted error rates, we also fit 2 AFM models for the Pre-tutor subset and Post-tutor subset. Later we compute the predicted error rate with these two models.

Figure 4 shows three predicted error rate series. The blue dash line represents the predicted error rate for the pre-tutor subset and the black dash line represents the predicted error rate for the post-tutor subset. The red curve represents the predicted error rate computed by our New AFM model. We can see that the predicted error rate decreases quickly in the beginning, but as opportunity increases, it would decrease at a much lower speed, suggesting that the learning rate has comparatively decreased.

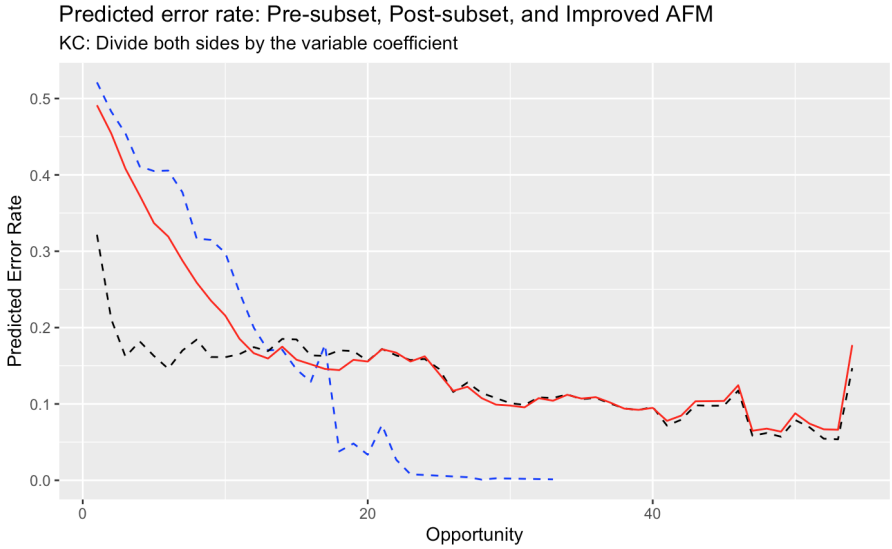


Figure 4: Predicted error rate for three AFM models for KC "Divide both sides by the variable coefficient"

Table 3 shows the summary of the new AFM model. As the coefficients of our Teacher-indicator and Opportunity are greater than 0, we identify a positive effect of tutor' intervention and students' natural learning process. However, we notice a negative coefficient for the interaction term, which suggests that the learning rate (which is defined as the coefficient of Opportunity) would decrease after the teacher intervened in the learning process, as the coefficient for Opportunity would decrease after Teacher-indicator switching from 0 to 1. This observation contradicts our assumption, as we assume tutor intervention would be effective in improving both the effect and the learning rate in a learning process. It is worth noting that this situation is consistent for all KCs. More detailed analysis can be found in Appendix 2-5.

Table 3. Estimated coefficients for KC "Divide both sides by the variable coefficient"

Variable	Coefficient	P-value
Intercept	0.16770	0.581
Teacher-indicator	2.19549	2.93×10^{-12}
Opportunity	0.19765	2.38×10^{-11}
Teacher-indicator*Opportunity	-0.18750	7.09×10^{-10}

4.1.2 KC: "Combine constant terms"

The second KC is "Divide both sides by the variable coefficient". We use the same procedure mentioned in section 4.1.1 and visualize the results. In Figure

5, the blue dash line represents the predicted error rate for the pre-tutor subset. The black dash line represents the predicted error rate for the Post-tutor subset, and the red curve represents the predicted error rate computed by our New AFM model. We observe a similar pattern as last KC that the predicted error rate for the New AFM model decreases quickly in the beginning but much lower later, suggesting that the learning rate has comparatively decreased.

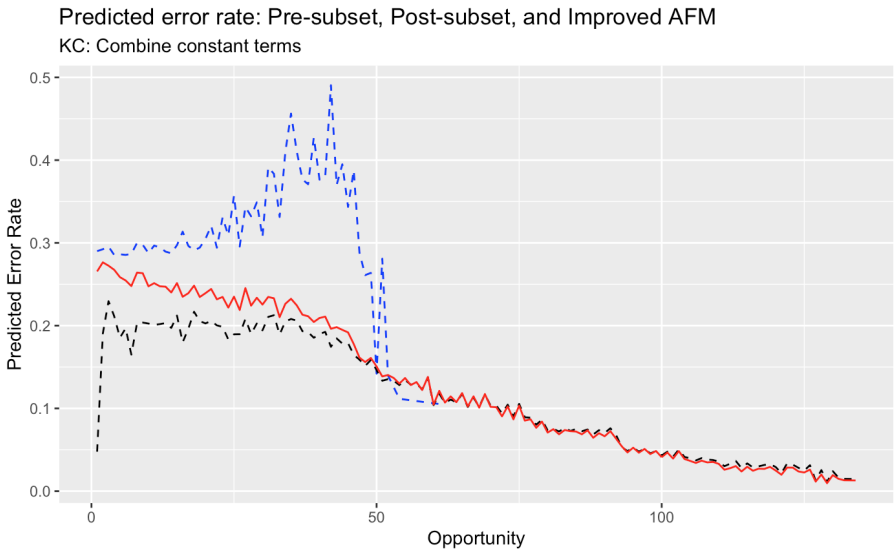


Figure 5: Predicted error rate for three AFM models for KC "Combine constant terms"

The results of the new AFM model on KC: Combine Constant Terms is shown in table 4. We also observe positive coefficients for Teacher-indicator and Opportunity and a negative coefficient for the interaction term, which also contradicts our assumption. However, none of these coefficients are statistically significant.

Table 4. Estimated coefficients for KC "Combine constant terms"

Variable	Coefficient	P-value
Intercept	1.41282	0.581
Teacher-indicator	0.20702	0.1117
Opportunity	0.01087	0.0632
Teacher-indicator*Opportunity	-0.00392	0.5139

4.1.3 Current Findings and Motivation for The Next Step

From the results so far, we've identified a consistent situation for the two KCs, which is the negative coefficient for the interaction term between Teacher-indicator and Opportunity. One possible guess is that students who get teachers'

intervention early might be systematically different from students who get the teachers' intervention late. If that is the case, as the opportunity increases, students in the pre-tutor subset would gradually switch to the post-teacher subset, leading to a great change in the sample size of these two groups. Eventually, all students would become post-tutor students. If there is a difference between early-intervened students and late-intervened students (for example, if students who receive tutor intervention early tend to have a lower error rate), it would cause a great change in the overall predicted error rate computed by our New AFM model, and thus influence our results. To verify our assumption, we utilize Method 2 to explore our assumptions.

4.2 Results from method 2

4.2.1 KC: "Divide both sides by the variable coefficient"

To verify our assumption in Result 1, we first explore the raw error rate for three groups in Figure 6. The black curve represents the error rate for students who received teachers' help in an early stage. The blue curve represents the error rate for the normal group, and the orange curve represents the error rate for the late group. We observe abnormality after opportunity 40, as the error rate increases, which is supposed to be decreasing as opportunity increases. We identify that the fluctuation in the raw error rate mainly comes from the sample size. Specifically, there were only 3 students left in the Late-intervention group after opportunity 40, 4 students left in the Normal-intervention group, and 1 student left in the Early-intervention group, while there were 15, 24, 30 students in each group at opportunity 10. To get rid of the impact of the imbalanced sample size, we drop data after opportunity 40 and use data before opportunity 40 to fit the Group AFM model and plot the predicted error rate for these three groups.

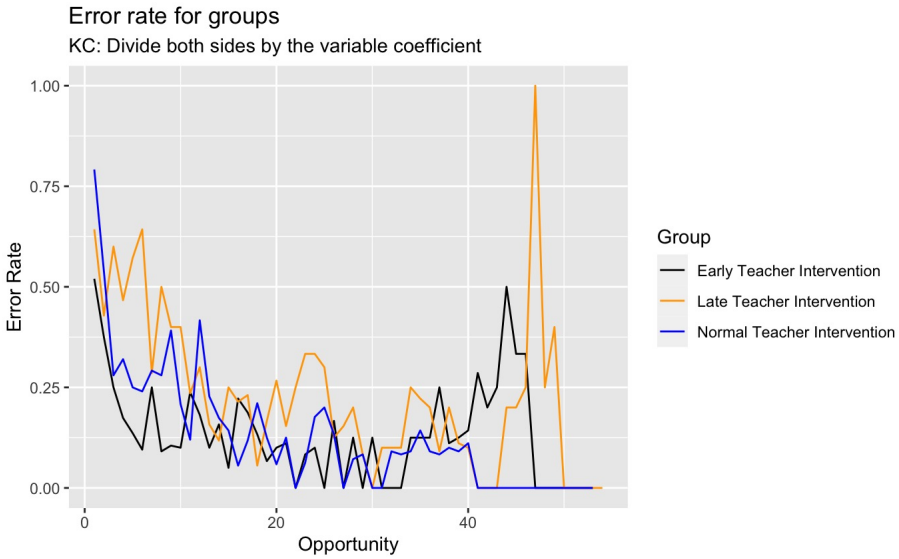


Figure 6: Error rate for three groups in KC "Divide both sides by the variable coefficient"

The predicted error rate is shown in Figure 7, and we notice that the Early-intervention group generally has the lowest error rate followed by the Normal-intervention and the Late-intervention. These results verify our assumption that there's a difference between students who receive intervention at different times.

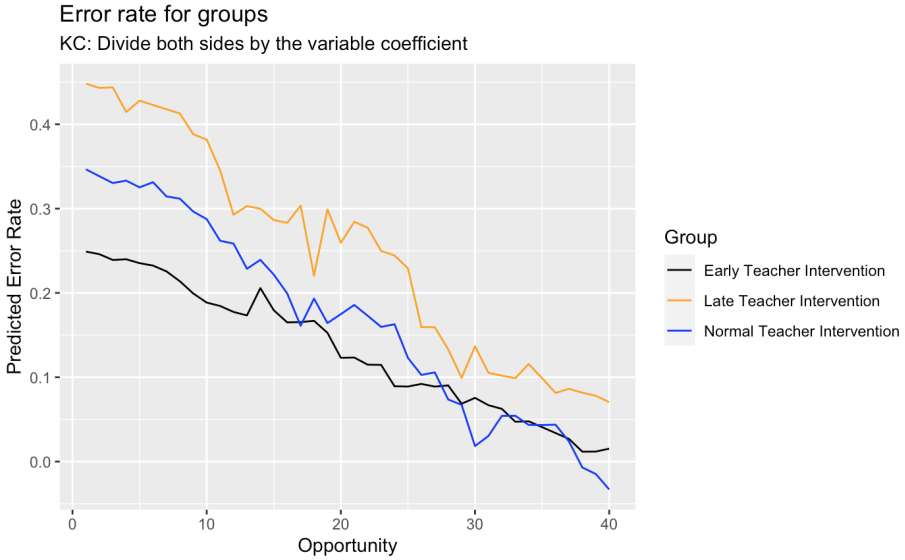


Figure 7: Predicted error rate for three groups in KC: "Divide both sides by the variable coefficient"

4.2.2 KC: "Combine constant terms"

For KC Combine Constant Terms, we also explore the raw error rate for three groups: students who got teachers' intervention in an early, normal and late stage in the second KC in Figure 8. The black curve represents the error rate for students who received teachers' help in an early stage. The blue curve represents error rate for the normal group, and the orange curve represents the error rate for the late group.

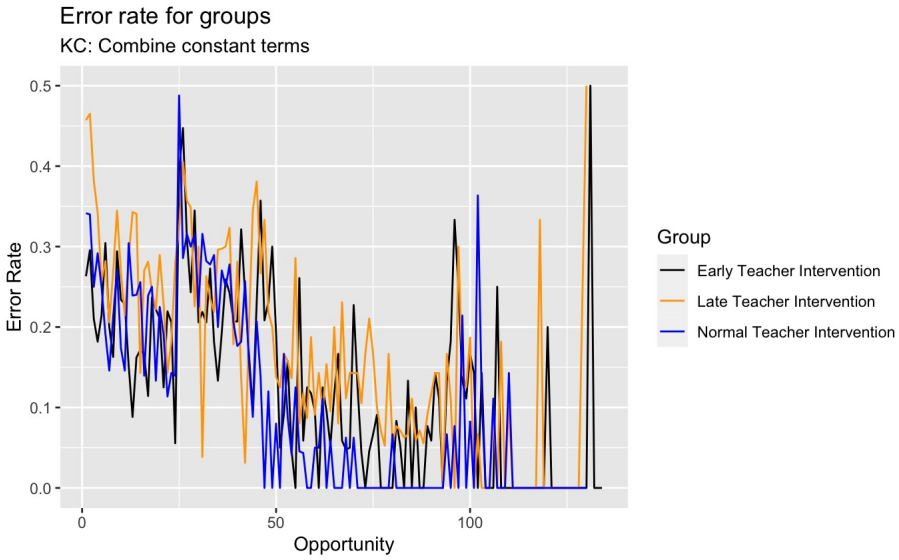


Figure 8: Error rate for three groups in KC "Combine constant terms"

After fitting the Group AFM model, we visualize the predicted error rate for these three groups in Figure 9. The results also suggests students who receive tutor intervention early tend to have generally lower error rate.

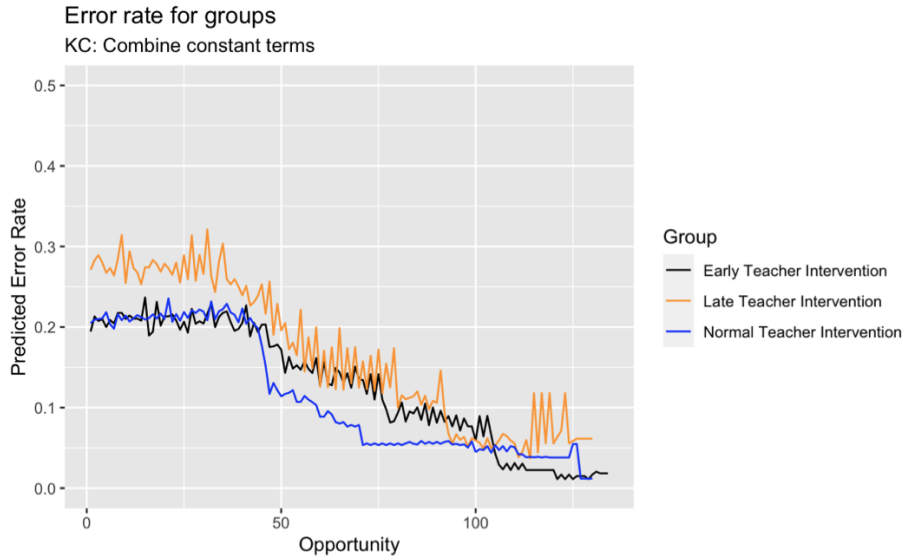


Figure 9: Predicted error rate for three groups in KC: "Combine constant terms"

5 Discussion

From the results of Method 1, we observe positive coefficients for teachers' intervention for both KC's. It suggests that teachers' intervention is effective at improving students' performances in terms of lowering the error rate. The effectiveness can be measured by the coefficients for Teacher-indicator and Teacher-Opportunity interaction term.

For both KC's we find that the coefficients of the Teacher-Opportunity interaction term are negative, indicating a negative effect in improving students' learning rate. This phenomenon is explained by our exploration in results from Method 2. We find that students who receive tutor' intervention late (i.e. in the Late-intervention group) would tend to perform poorly in general. One potential interpretation is that during students' learning process, all students who take questions related to a specific KC would eventually get teachers' intervention. In other words, more and more students would switch from the pre-tutor subset to the post-tutor subset as the opportunity increases. Also, due to the fact that students receiving teachers' intervention in a late-stage would systematically perform poorly in the questions, comparatively, these students switching from Pre-subset to Post-subset would highly influence our predicted error rate. It is not that the teachers would cause a negative effect on the learning process. It's just that students who tend to perform comparatively poorly in this KC gradually switch from the pre-subset to the post-subset, and influence our model. Thus, we would still consider the tutor's intervention to be effective.

In future studies, we would recommend starting from the following three aspects. First, we can try the new AFM model on another dataset. We have implemented this model for 195 students in a math tutor class. It would be worth checking the consistency of the model by implementing it in different sessions or other classes. Second, exploring the relationship between the number of teachers' intervention with students' learning rate could help strengthen our study. As many students have received teachers' help more than once, It might be interesting to check whether the number of interventions influences students' performance. However, in our study, we only consider the first intervention time in our analysis. Finally, we should examine the difference in learning rates between students who received teachers' intervention and students who did not, using the AFM model.

6 Reference

[1]Ido Roll, Ryan S. J. d. Baker, Vincent Aleven Kenneth R. Koedinger (2014) On the Benefits of Seeking (and Avoiding) Help in Online Problem-Solving Environments, Journal of the Learning Sciences, 23:4, 537-560, DOI: 10.1080/10508406.2014.883977

[2]Koedinger, K.R., Baker, R.S.J.d., Cunningham, K., Skogsholm, A., Leber, B., Stamper, J. (2010) A Data Repository for the EDM community: The PSLC DataShop. In Romero, C., Ventura, S., Pechenizkiy, M., Baker, R.S.J.d. (Eds.) Handbook of Educational Data Mining. Boca Raton, FL: CRC Press.

[3]Aleven, V., Koedinger, K. R. (2013). Knowledge component approaches to learner modeling. In R. Sottilare, A. Graesser, X. Hu, H. Holden (Eds.), Design recommendations for adaptive intelligent tutoring systems (Vol. I, Learner Modeling, pp. 165-182). Orlando, FL: US Army Research Laboratory.

[4]Pavlik Jr, Phil Cen, Hao Koedinger, Kenneth. (2009). Performance Factors Analysis - A New Alternative to Knowledge Tracing. Frontiers in Artificial Intelligence and Applications. 200. 531-538. 10.3233/978-1-60750-028-5-531.

[5]Cen H., Koedinger K., Junker B. (2008) Comparing Two IRT Models for Conjunctive Skills. In: Woolf B.P., Aïmeur E., Nkambou R., Lajoie S. (eds) Intelligent Tutoring Systems. ITS 2008. Lecture Notes in Computer Science, vol 5091. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-69132-$7_1$11

Appendices: Detecting Learning Discontinuity for Out-of-tutor Event

Yiwen Zhang, Naifei Pan, Jie Luo

05/15/2021

Contents


```
library(tidyverse)
library(ggpubr)
```

Appendix 1: Initial Data Import & Exploration

Load data

The loaded data is combined version of transaction dataset and studentstep data set which we added the tutor intervention time. Due to the run-time limitation, we will show the R code of bridging the data along with our final products.

```
student_step <- read.delim("student_step.txt")
HCI <- read.csv('HCI_final.csv')
```

We have 195 students in total

```
length(unique(student_step$Anon.Student.Id))
```

```
## [1] 195
```

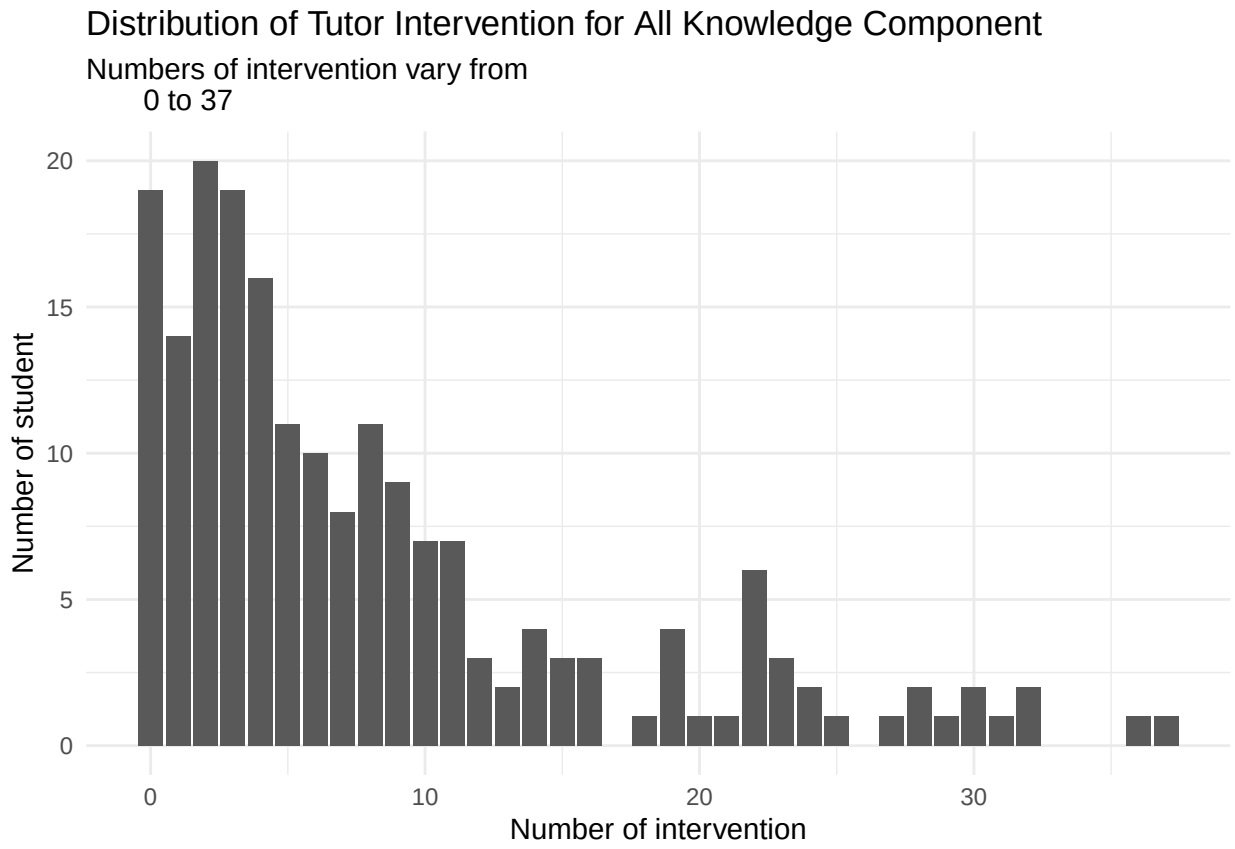
In general, we have 7 KC.

```
unique(student_step$KC..Default.)[1:8]
```

```
## [1] "Add/subtract_constant_from_both_sides"
## [2] ""
## [3] "Combine_constant_terms"
## [4] "Divide_both_sides_by_the_variable_coefficient"
## [5] "Compute_quotient_for_constant"
## [6] "Compute_quotient_for_variable_coefficient"
## [7] "Add/subtract_variable_from_both_sides"
## [8] "Combine_variable_terms"
```

Distribution of the number of intervention

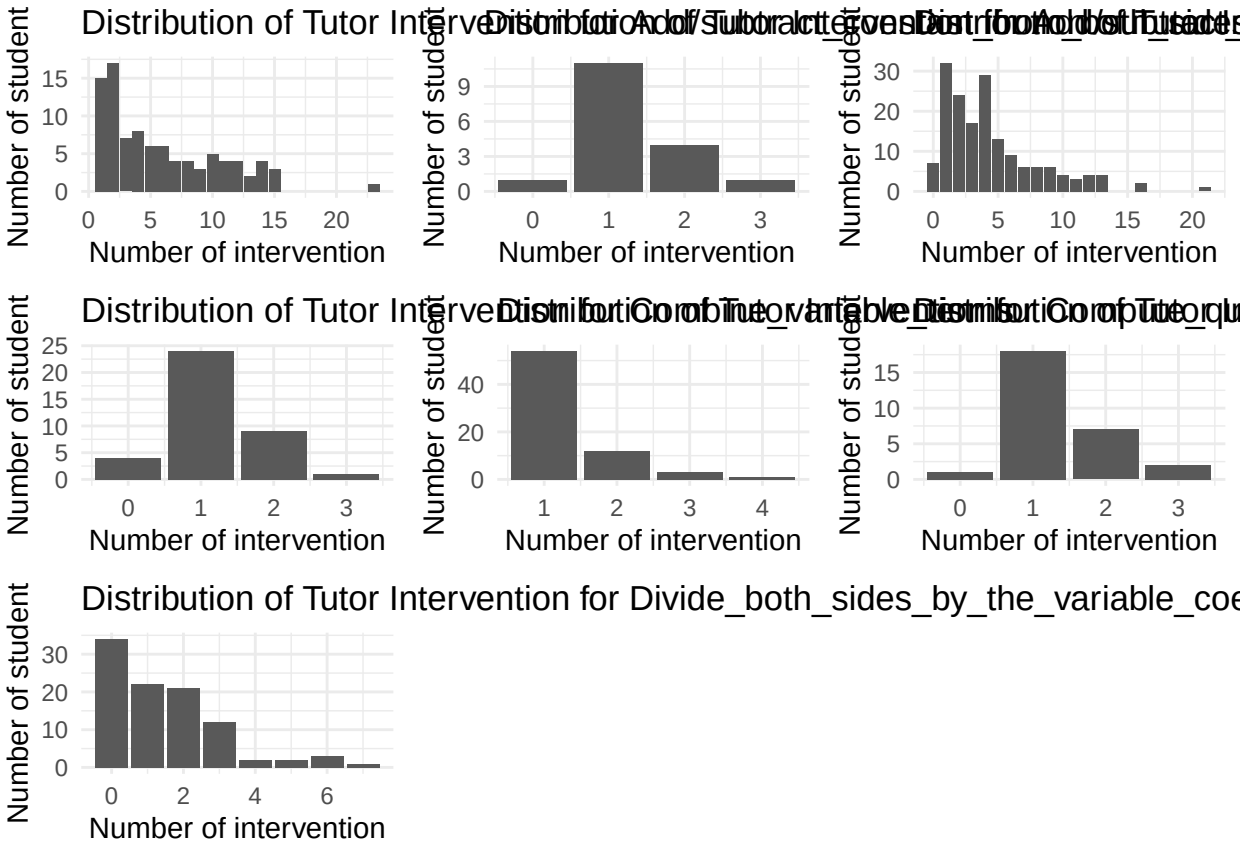
```
tutor_num <- HCI %>%
  select(Anon.Student.Id, IfTutor) %>%
  group_by(Anon.Student.Id) %>%
  summarise(num_intervention = sum(IfTutor)) %>%
  arrange(num_intervention)

num_tutor <- tutor_num %>%
  select(num_intervention, Anon.Student.Id) %>%
  group_by(num_intervention) %>%
  summarise(number_student = n()) %>%
  add_row(num_intervention=0, number_student=18, .before = 1)

ggplot(data=num_tutor, aes(x=num_intervention, y= number_student))+
  geom_bar(stat="identity") +
  labs(title="Distribution of Tutor Intervention for All Knowledge Component" ,
       y="Number of student",
       x= "Number of intervention", subtitle = "Numbers of intervention vary from
       0 to 37")+
  theme_minimal()
```



Distribution for each KC

```
kc <- c("Add/subtract_constant_from_both_sides",
        "Add/subtract_variable_from_both_sides",
        "Combine_constant_terms",
        "Combine_variable_terms",
        "Compute_quotient_for_constant",
        "Compute_quotient_for_variable_coefficient",
        "Divide_both_sides_by_the_variable_coefficient")

for(k in 1:7){
  df <- HCI %>%
    filter(grepl(kc[k], KC..Default.)) %>%
    select(Anon.Student.Id, IfTutor) %>%
    group_by(Anon.Student.Id) %>%
    summarise(num_intervention = sum(IfTutor)) %>%
    arrange(num_intervention)   %>%
    select(num_intervention, Anon.Student.Id) %>%
    group_by(num_intervention) %>%
    summarise(number_student = n())

    assign(paste0("kc_",k), ggplot(data=df, aes(x=num_intervention,
                                          y= number_student))+
                        geom_bar(stat="identity") +
                        labs(title=paste("Distribution of Tutor Intervention for",
                                         kc[k]),
                             y="Number of student",
```

```
                          x= "Number of intervention")+
                      theme_minimal())
}

ggarrange(kc_1,kc_2,kc_3,kc_4,kc_5,kc_6,kc_7)
```



Appendix 2: Analysis of The First KC & New AFM model

KC:Divide both sides by the variable coefficient

For the analysis, We observed the similar pattern and results for all KC. In this report, we picked two KCs for the analysis.

```
HCI1 <- HCI %>%
  filter(grepl('Divide_both_sides_by_the_variable_coefficient', KC..Default.))
```

97 students received the tutor helps for this KC

```
length(unique(HCI1$Anon.Student.Id))
```

```
## [1] 97
```

Split the opportunity with situation like: "KC:Add/subtract_constant_from_both_sides ~~ Combine_constant_terms Opportunity: 10 ~~ 19"

```
kcname <- "Divide_both_sides_by_the_variable_coefficient"
kc_character <- as.character(HCI1$KC..Default.)
Oppo_character <- as.character(HCI1$Opportunity..Default.)
rows_multi <- which(grepl('~~', Oppo_character))
oppo_first <- rep(0, length(rows_multi))

for (i in 1:length(rows_multi)) {
  str_kc<- unlist(strsplit(kc_character[rows_multi[i]], split = '~~'))
  str_oppo <- unlist(strsplit(Oppo_character[rows_multi[i]], split = '~~'))
  if(str_kc[1] == kcname){
    oppo_first[i] <- str_oppo[1]
  }
  if(str_kc[2] == kcname){
    oppo_first[i] <- str_oppo[2]
  }
}

Oppo_character[rows_multi] <- oppo_first
# Opportunity we will be using
HCI1$Opportunity_Numeric <- as.numeric(Oppo_character)
```

Indicator variable suggesting pre-post tutor observation.

```
tutor.indicator <- rep(0, nrow(HCI1))
for (i in 1:nrow(HCI1)){
  if(as.numeric(as.character(HCI1$Opportunity_Numeric[i])) <= HCI1$TutorTime[i])
    tutor.indicator[i] <- 0
  if(as.numeric(as.character(HCI1$Opportunity_Numeric[i])) > HCI1$TutorTime[i])
    tutor.indicator[i] <- 1
}
HCI1$Post <- tutor.indicator
```

Calculate the original error rate(true error rate) Using success variable

```
L1 = length(HCI1$Anon.Student.Id)
Success1 = vector(mode="numeric", length=L1)
Success1[HCI1$First.Attempt=="correct"]=1
HCI1$Success1 <- Success1
true_rate <- HCI1 %>%
  select(Opportunity_Numeric, Success1) %>%
  group_by(Opportunity_Numeric) %>%
  summarise("Students with help" = 1- sum(Success1)/n())
```

Find the students that did not receive the help in this KC

```
full_name <- unique(student_step$Anon.Student.Id)

name_177 <- unique(HCI1$Anon.Student.Id)

no_help <- full_name[!(full_name %in% name_177)]

stu_no_help <- student_step %>%
```

```
  filter(Anon.Student.Id %in% no_help) %>%
  filter(grepl('Divide_both_sides_by_the_variable_coefficient', KC..Default.))
```

There are 84 students who didn't received any helps in this KC

```
length(unique(stu_no_help$Anon.Student.Id))
```

```
## [1] 84
```

Processing the data for students that did not receive the helps

```
kcname <- "Divide_both_sides_by_the_variable_coefficient"
kc_character <- as.character(stu_no_help$KC..Default.)
Oppo_character <- as.character(stu_no_help$Opportunity..Default.)
rows_multi <- which(grepl('~~', Oppo_character))
oppo_first <- rep(0, length(rows_multi))

for (i in 1:length(rows_multi)) {
  str_kc<- unlist(strsplit(kc_character[rows_multi[i]], split = '~~'))
  str_oppo <- unlist(strsplit(Oppo_character[rows_multi[i]], split = '~~'))
  if(str_kc[1] == kcname){
    oppo_first[i] <- str_oppo[1]
  }
  if(str_kc[2] == kcname){
    oppo_first[i] <- str_oppo[2]
  }
}

Oppo_character[rows_multi] <- oppo_first
# Opportunity we will be using
stu_no_help$Opportunity_Numeric <- as.numeric(Oppo_character)
```

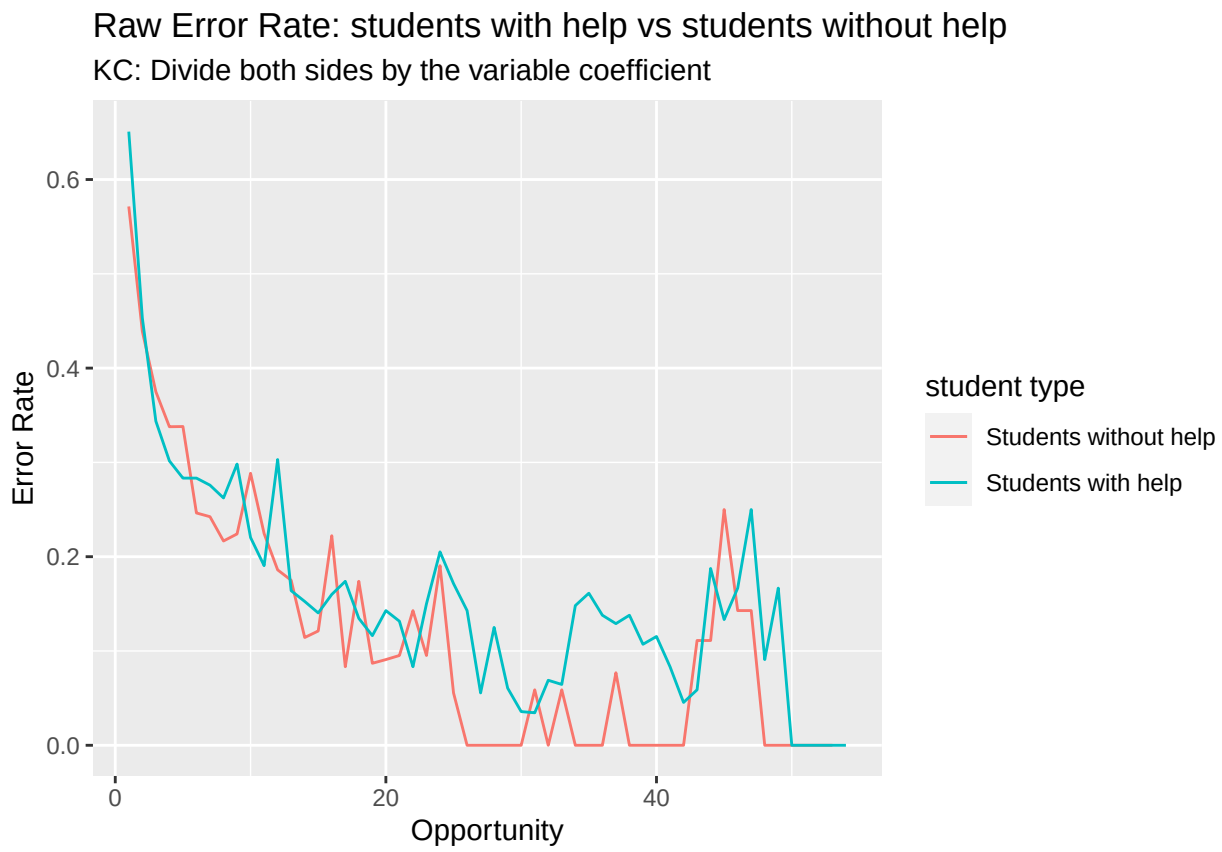
Error rate with student with no helps using success variable

```
L2 = length(stu_no_help$Anon.Student.Id)
Success2 = vector(mode="numeric", length=L2)
Success2[stu_no_help$First.Attempt=="correct"]=1
stu_no_help$Success2 <- Success2
true_rate_no_help <- stu_no_help %>%
  select(Opportunity_Numeric, Success2) %>%
  group_by(Opportunity_Numeric) %>%
  summarise("Students without help" = 1- sum(Success2)/n())
```

plot raw error rate students with no help vs student with tutor

```
ggplot() +
  geom_line(data=true_rate_no_help, aes(x= Opportunity_Numeric,
                                        y =`Students without help`,
                                        color="balck"))+
  geom_line(data=true_rate, aes(x=Opportunity_Numeric,
                                y = `Students with help`,color="Blue"))+
```

```
  labs(title="Raw Error Rate: students with help vs students without help",
       subtitle="KC: Divide both sides by the variable coefficient",
       x="Opportunity",
       y="Error Rate")+
  scale_color_discrete(name="student type", labels = c("Students without help" ,
                                                       "Students with help" ))
```



Modeling

```
library(lme4)
AFM1 <- glmer(Success1 ~ (1|Anon.Student.Id) + Post + Opportunity_Numeric +
                Opportunity_Numeric:Post, family=binomial(), data= HCI1)

summary(AFM1)
```

```
## Generalized linear mixed model fit by maximum likelihood (Laplace
##   Approximation) [glmerMod]
##  Family: binomial  ( logit )
## Formula: Success1 ~ (1 | Anon.Student.Id) + Post + Opportunity_Numeric +
##     Opportunity_Numeric:Post
##    Data: HCI1
##
##      AIC      BIC   logLik deviance df.resid
##   1361.7   1389.6   -675.8   1351.7     1949
##
```

```
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -6.8242  0.0934  0.1593  0.3460  3.0095
##
## Random effects:
##  Groups          Name        Variance Std.Dev.
##  Anon.Student.Id (Intercept) 3.614    1.901
## Number of obs: 1954, groups:  Anon.Student.Id, 97
##
## Fixed effects:
##                            Estimate Std. Error z value Pr(>|z|)
## (Intercept)                 0.16770    0.30406   0.552    0.581
## Post                        2.19549    0.31450   6.981 2.93e-12 ***
## Opportunity_Numeric         0.19765    0.02959   6.680 2.38e-11 ***
## Post:Opportunity_Numeric -0.18750    0.03042  -6.164 7.10e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##             (Intr) Post   Oppr_N
## Post        -0.409
## Opprtnty_Nm -0.515  0.482
## Pst:Opprt_N  0.484 -0.675 -0.947
```

Prediction rate

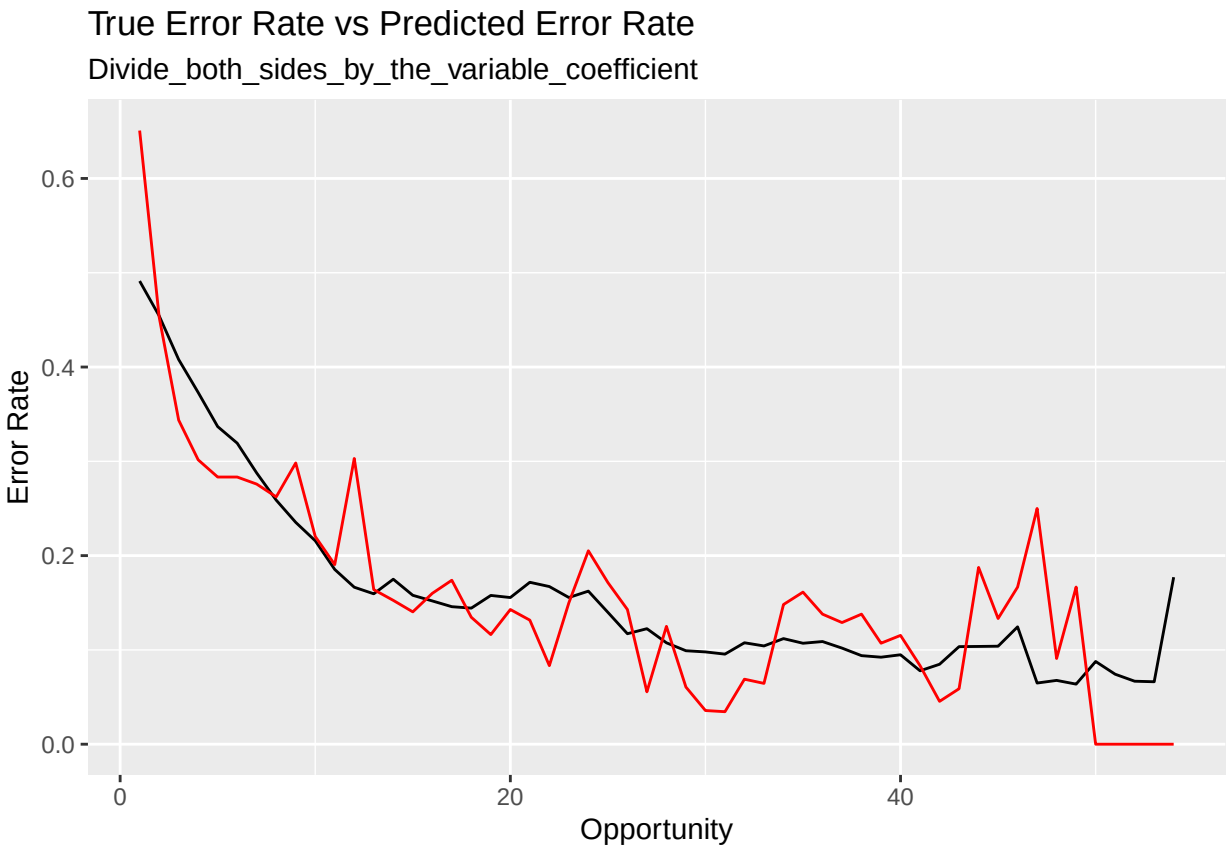
```
pred1 <- predict(AFM1, HCI1, type="response")
HCI1$Pred <- pred1

df1 <- data.frame(Opportunity = HCI1$Opportunity_Numeric, Pred = HCI1$Pred) %>%
  group_by(Opportunity) %>%
  summarise(error= 1 - mean(Pred))
```

Ture error rate

```
HCI1$Success1 <- Success1
true_rate <- HCI1 %>%
  select(Opportunity_Numeric, Success1) %>%
  group_by(Opportunity_Numeric) %>%
  summarise(error = 1- sum(Success1)/n())
```

Plot the prediction result vs the true rate

```
ggplot() +
  geom_line(data=df1, aes(x= Opportunity, y = error), col="Black")+
  geom_line(data=true_rate, aes(x= Opportunity_Numeric, y = error),col="red")+
  labs(title="True Error Rate vs Predicted Error Rate",subtitle=kcname,
       x="Opportunity", y="Error Rate")
```



Appendix 3: AFM Mdoels of Pre & Post Tutor Intervention for The First KC

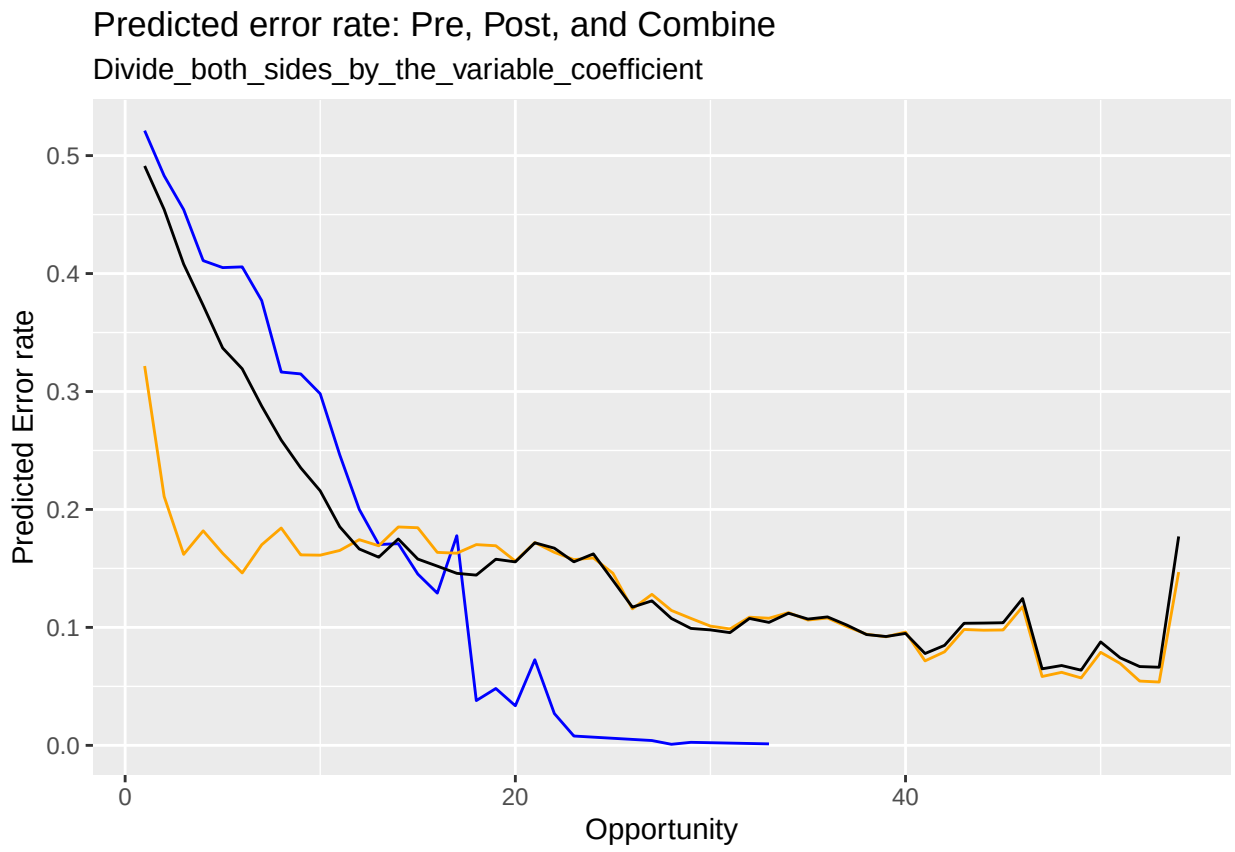
Fit 2 individual AFMs for pre and post subset

```
HCI1_pre <- HCI1 %>%
  filter(TutorTime > Opportunity_Numeric)
HCI1_post <- HCI1 %>%
  filter(TutorTime <= Opportunity_Numeric)

L1_pre = length(HCI1_pre$Anon.Student.Id)
Success1_pre = vector(mode="numeric", length=L1_pre)
Success1_pre[HCI1_pre$First.Attempt=="correct"]=1
AFM1_pre <- glmer(Success1_pre ~ (1|Anon.Student.Id) + Opportunity_Numeric,
                  family=binomial(), data= HCI1_pre)
pred1_pre <- predict(AFM1_pre, HCI1_pre, type="response")
HCI1_pre$Pred <- pred1_pre

L1_post = length(HCI1_post$Anon.Student.Id)
Success1_post = vector(mode="numeric", length=L1_post)
Success1_post[HCI1_post$First.Attempt=="correct"]=1
AFM1_post <- glmer(Success1_post ~ (1|Anon.Student.Id) + Opportunity_Numeric,
                   family=binomial(), data= HCI1_post)
pred1_post <- predict(AFM1_post, HCI1_post, type="response")
HCI1_post$Pred <- pred1_post
```

```
df1_pre <- data.frame(Opportunity = HCI1_pre$Opportunity_Numeric,
                      Pred = HCI1_pre$Pred) %>%
  group_by(Opportunity) %>%
  summarise(error= 1 - mean(Pred))

df1_post <- data.frame(Opportunity = HCI1_post$Opportunity_Numeric,
                       Pred = HCI1_post$Pred) %>%
  group_by(Opportunity) %>%
  summarise(error= 1 - mean(Pred))
```

Visualize the results

```
ggplot()+
  geom_line(data=df1_pre, aes(x=Opportunity, y=error),col="Blue") +
  geom_line(data=df1_post, aes(x=Opportunity, y=error), col="Orange")+
  geom_line(data=df1, aes(x= Opportunity, y = error), col="Black")+
  labs(title="Predicted error rate: Pre, Post, and Combine",subtitle=kcname,
       x="Opportunity", y="Predicted Error rate")
```



Appendix 4: Analysis of The Second KC & New AFM Model

KC:Combine constant terms

```
HCI2 <- HCI %>%
  filter(grepl('Combine_constant_terms', KC..Default.))
```

167 students received the tutor helps for this KC

```
length(unique(HCI2$Anon.Student.Id))
```

```
## [1] 167
```

Split the opportunity with situation like: "KC:Add/subtract_constant_from_both_sides ~~ Combine_constant_terms Opportunity: 10 ~~ 19"

```
kcname <- "Combine_constant_terms"
kc_character <- as.character(HCI2$KC..Default.)
Oppo_character <- as.character(HCI2$Opportunity..Default.)
rows_multi <- which(grepl('~~', Oppo_character))
oppo_first <- rep(0, length(rows_multi))

for (i in 1:length(rows_multi)) {
  str_kc<- unlist(strsplit(kc_character[rows_multi[i]], split = '~~'))
  str_oppo <- unlist(strsplit(Oppo_character[rows_multi[i]], split = '~~'))
  if(str_kc[1] == kcname){
    oppo_first[i] <- str_oppo[1]
  }
  if(str_kc[2] == kcname){
    oppo_first[i] <- str_oppo[2]
  }
}

Oppo_character[rows_multi] <- oppo_first
# Opportunity we will be using
HCI2$Opportunity_Numeric <- as.numeric(Oppo_character)
```

Indicator variable suggesting pre-post tutor observation.

```
tutor.indicator <- rep(0, nrow(HCI2))
for (i in 1:nrow(HCI1)){
  if(as.numeric(as.character(HCI2$Opportunity_Numeric[i])) <= HCI2$TutorTime[i])
    tutor.indicator[i] <- 0
  if(as.numeric(as.character(HCI2$Opportunity_Numeric[i])) > HCI2$TutorTime[i])
    tutor.indicator[i] <- 1
}
HCI2$Post <- tutor.indicator
```

Calculate the original error rate(true error rate) Using success variable

```
L1 = length(HCI2$Anon.Student.Id)
Success1 = vector(mode="numeric", length=L1)
```

```
Success1[HCI2$First.Attempt=="correct"]=1
HCI2$Success1 <- Success1
true_rate <- HCI2 %>%
  select(Opportunity_Numeric, Success1) %>%
  group_by(Opportunity_Numeric) %>%
  summarise("Students with help" = 1- sum(Success1)/n())
```

Find the students that did not receive the help in this KC

```
full_name <- unique(student_step$Anon.Student.Id)

name_177 <- unique(HCI2$Anon.Student.Id)

no_help <- full_name[!(full_name %in% name_177)]

stu_no_help <- student_step %>%
  filter(Anon.Student.Id %in% no_help) %>%
  filter(grepl('Combine_constant_terms', KC..Default.))
```

There are 27 students who didn't received any helps in this KC

```
length(unique(stu_no_help$Anon.Student.Id))
```

```
## [1] 27
```

Processing the data for students that did not receive the helps

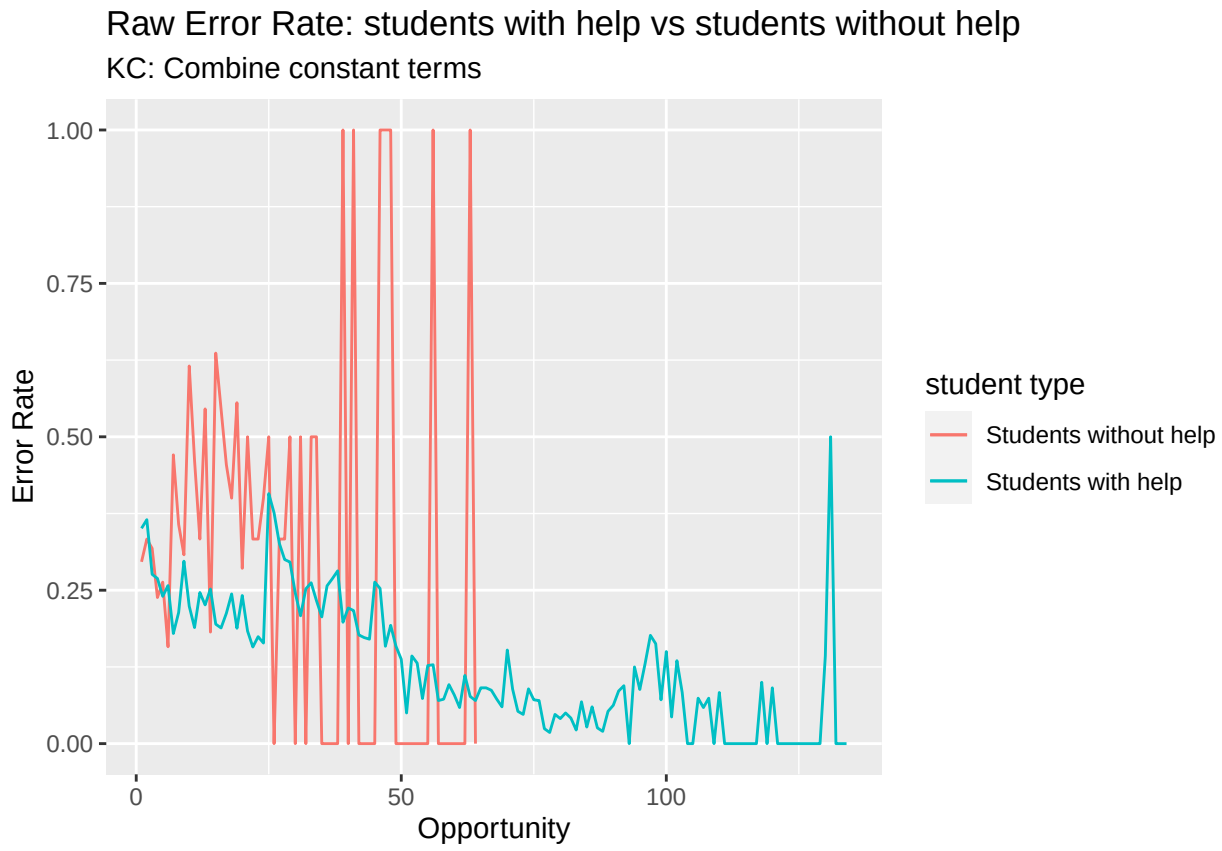
```
kcname <- "Combine_constant_terms"
kc_character <- as.character(stu_no_help$KC..Default.)
Oppo_character <- as.character(stu_no_help$Opportunity..Default.)
rows_multi <- which(grepl('~~', Oppo_character))
oppo_first <- rep(0, length(rows_multi))

for (i in 1:length(rows_multi)) {
  str_kc<- unlist(strsplit(kc_character[rows_multi[i]], split = '~~'))
  str_oppo <- unlist(strsplit(Oppo_character[rows_multi[i]], split = '~~'))
  if(str_kc[1] == kcname){
    oppo_first[i] <- str_oppo[1]
  }
  if(str_kc[2] == kcname){
    oppo_first[i] <- str_oppo[2]
  }
}

Oppo_character[rows_multi] <- oppo_first
# Opportunity we will be using
stu_no_help$Opportunity_Numeric <- as.numeric(Oppo_character)
```

Error rate with student with no helps using success variable

```
L2 = length(stu_no_help$Anon.Student.Id)
Success2 = vector(mode="numeric", length=L2)
Success2[stu_no_help$First.Attempt=="correct"]=1
stu_no_help$Success2 <- Success2
true_rate_no_help <- stu_no_help %>%
  select(Opportunity_Numeric, Success2) %>%
  group_by(Opportunity_Numeric) %>%
  summarise("Students without help" = 1- sum(Success2)/n())
```

plot raw error rate students with no help vs student with tutor

```
ggplot() +
  geom_line(data=true_rate_no_help, aes(x= Opportunity_Numeric,
                                        y =`Students without help`,
                                        color="balck"))+
  geom_line(data=true_rate, aes(x=Opportunity_Numeric,
                                y = `Students with help`,color="Blue"))+
  labs(title="Raw Error Rate: students with help vs students without help",
       subtitle="KC: Combine constant terms", x="Opportunity", y="Error Rate")+
  scale_color_discrete(name="student type", labels = c("Students without help" ,
                                                       "Students with help" ))
```



Modeling

```
AFM2 <- glmer(Success1 ~ (1|Anon.Student.Id) + Post + Opportunity_Numeric +
                Opportunity_Numeric:Post, family=binomial(), data= HCI2)

summary(AFM2)
```

```
## Generalized linear mixed model fit by maximum likelihood (Laplace
##   Approximation) [glmerMod]
##  Family: binomial  ( logit )
## Formula: Success1 ~ (1 | Anon.Student.Id) + Post + Opportunity_Numeric +
##     Opportunity_Numeric:Post
##    Data: HCI2
##
##      AIC      BIC   logLik deviance df.resid
##   6106.5   6141.6  -3048.3   6096.5     8280
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -9.5438  0.1218  0.2203  0.4026  2.0364
##
## Random effects:
##  Groups          Name        Variance Std.Dev.
##  Anon.Student.Id (Intercept) 2.049    1.431
## Number of obs: 8285, groups:  Anon.Student.Id, 167
##
## Fixed effects:
##                            Estimate Std. Error z value Pr(>|z|)
## (Intercept)                1.527374   0.132499  11.527  < 2e-16 ***
## Post                      -0.276141   0.217470  -1.270    0.204
## Opportunity_Numeric        0.008647   0.001716   5.040 4.66e-07 ***
## Post:Opportunity_Numeric   0.004195   0.004360   0.962    0.336
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##             (Intr) Post   Oppr_N
## Post        -0.143
## Opprtnty_Nm -0.363  0.101
## Pst:Opprt_N  0.128 -0.713 -0.375
## convergence code: 0
## Model is nearly unidentifiable: very large eigenvalue
##  - Rescale variables?
```

Prediction rate

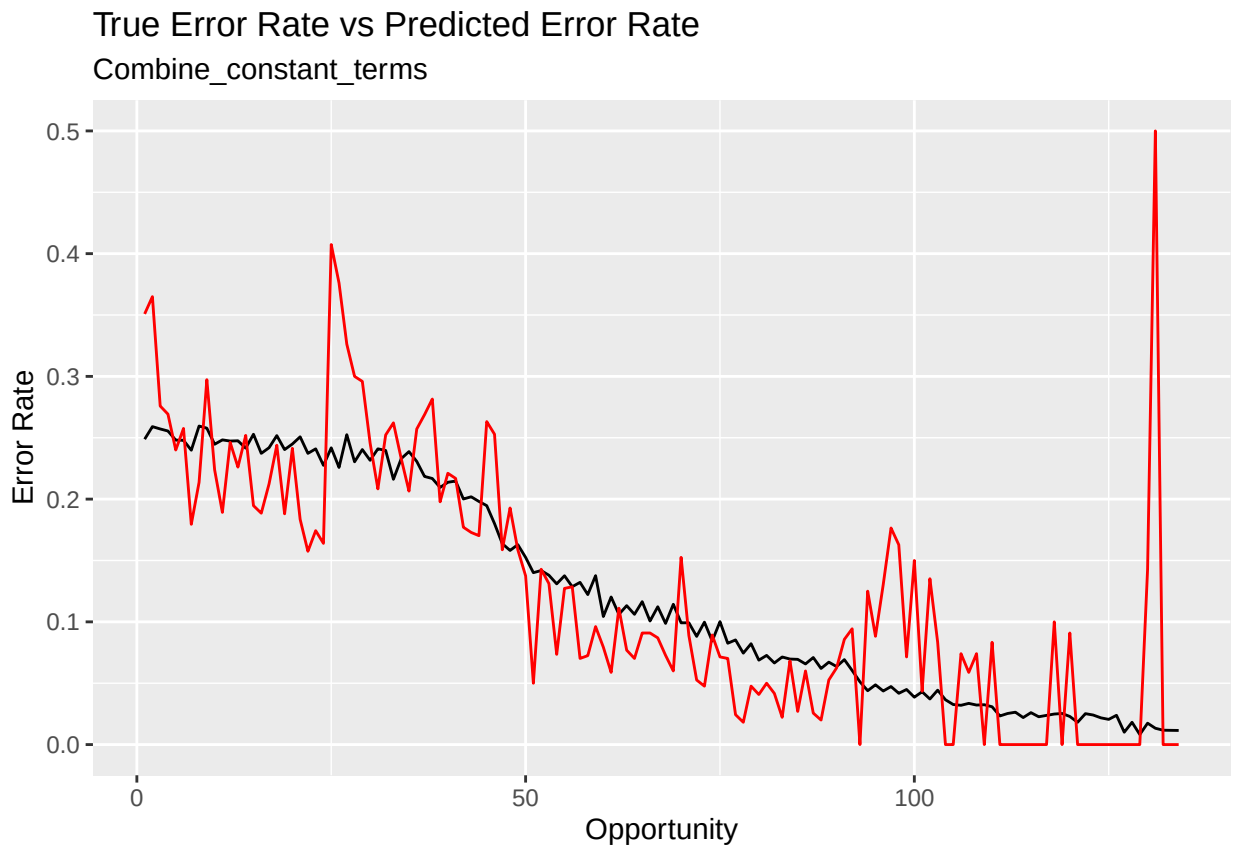
```
pred2 <- predict(AFM2, HCI2, type="response")
HCI2$Pred <- pred2

df2 <- data.frame(Opportunity = HCI2$Opportunity_Numeric, Pred = HCI2$Pred) %>%
  group_by(Opportunity) %>%
  summarise(error= 1 - mean(Pred))
```

True error rate

```
HCI2$Success1 <- Success1
true_rate <- HCI2 %>%
  select(Opportunity_Numeric, Success1) %>%
  group_by(Opportunity_Numeric) %>%
  summarise(error = 1- sum(Success1)/n())
```

Plot the prediction result vs the true rate

```
ggplot() +
  geom_line(data=df2, aes(x= Opportunity, y = error), col="Black")+
  geom_line(data=true_rate, aes(x= Opportunity_Numeric, y = error),col="red")+
  labs(title="True Error Rate vs Predicted Error Rate",subtitle=kcname,
       x="Opportunity", y="Error Rate")
```



Appendix 5: AFM Mdoels of Pre & Post Tutor Intervention for The Second KC

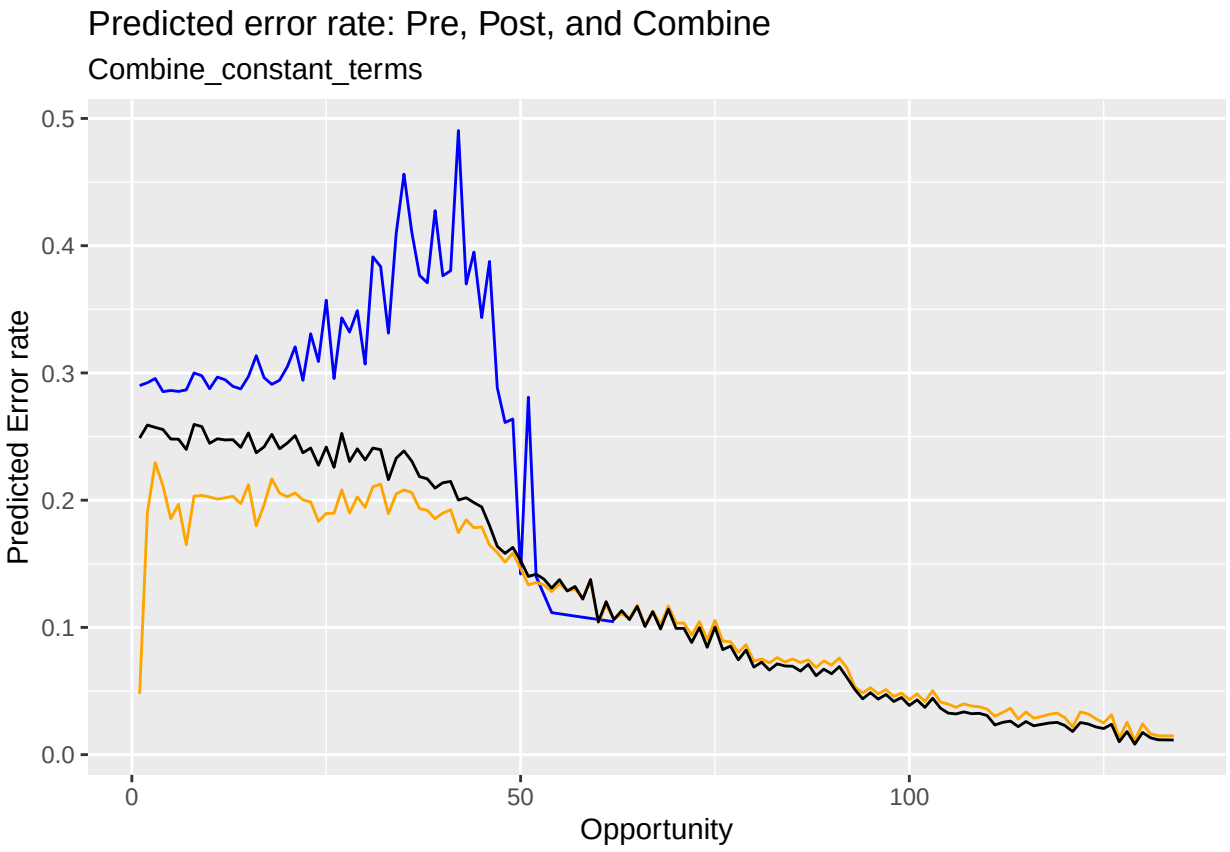
Fit 2 individual AFMs for pre and post subset

```
HCI2_pre <- HCI2 %>%
  filter(TutorTime > Opportunity_Numeric)
HCI2_post <- HCI2 %>%
  filter(TutorTime <= Opportunity_Numeric)

L2_pre = length(HCI2_pre$Anon.Student.Id)
```

```
Success2_pre = vector(mode="numeric", length=L2_pre)
Success2_pre[HCI2_pre$First.Attempt=="correct"]=1
AFM2_pre <- glmer(Success2_pre ~ (1|Anon.Student.Id) + Opportunity_Numeric,
                  family=binomial(), data= HCI2_pre)
pred2_pre <- predict(AFM2_pre, HCI2_pre, type="response")
HCI2_pre$Pred <- pred2_pre

L2_post = length(HCI2_post$Anon.Student.Id)
Success2_post = vector(mode="numeric", length=L2_post)
Success2_post[HCI2_post$First.Attempt=="correct"]=1
AFM2_post <- glmer(Success2_post ~ (1|Anon.Student.Id) + Opportunity_Numeric,
                   family=binomial(), data= HCI2_post)
pred2_post <- predict(AFM2_post, HCI2_post, type="response")
HCI2_post$Pred <- pred2_post

df2_pre <- data.frame(Opportunity = HCI2_pre$Opportunity_Numeric,
                      Pred = HCI2_pre$Pred) %>%
  group_by(Opportunity) %>%
  summarise(error= 1 - mean(Pred))

df2_post <- data.frame(Opportunity = HCI2_post$Opportunity_Numeric,
                       Pred = HCI2_post$Pred) %>%
  group_by(Opportunity) %>%
  summarise(error= 1 - mean(Pred))
```

Visualize the results

```
ggplot()+
  geom_line(data=df2_pre, aes(x=Opportunity, y=error),col="Blue") +
  geom_line(data=df2_post, aes(x=Opportunity, y=error), col="Orange")+
  geom_line(data=df2, aes(x= Opportunity, y = error), col="Black")+
  labs(title="Predicted error rate: Pre, Post, and Combine",subtitle=kcname,
       x="Opportunity", y="Predicted Error rate")
```



Appendix 6: Eerly, Normal, and Late Tutor Intervention Analysis for Both KC

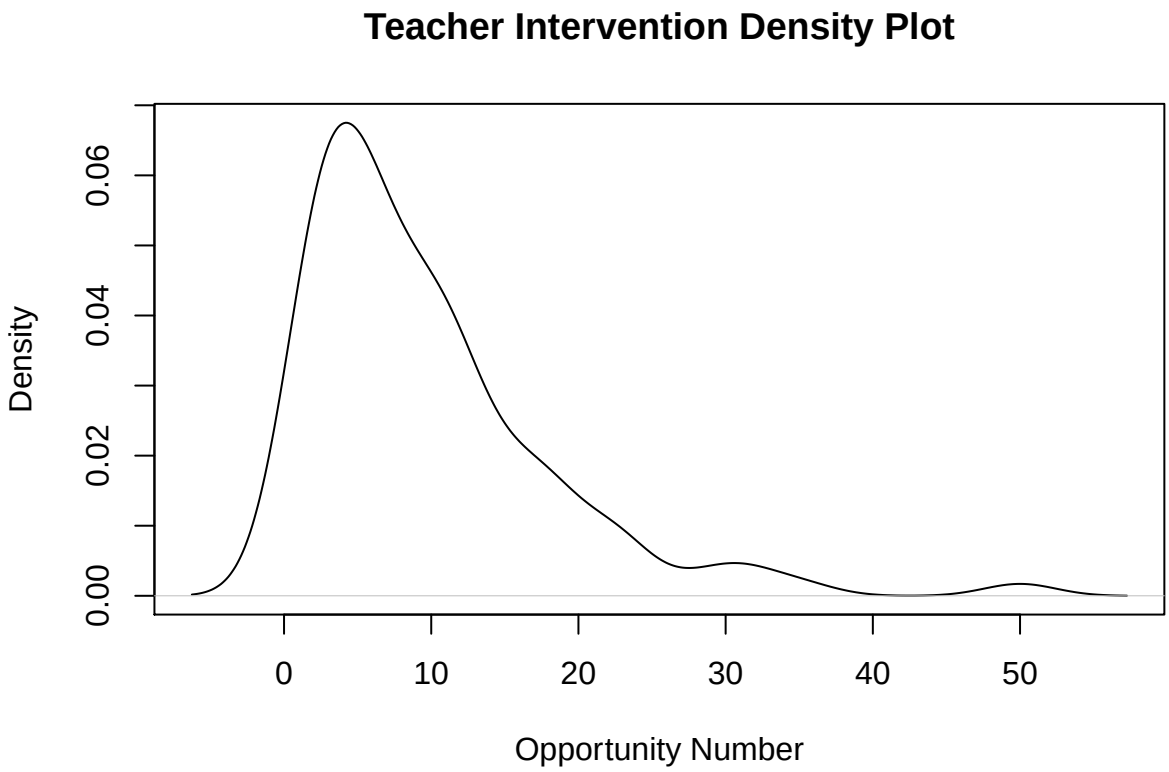
KC:Divide both sides by the variable coefficient

Divide students into early-tutor, normal-tutor, late-tutor groups.

```
library(data.table)
student_error <- data.table(Anon.Student.Id=character(),opportunity=numeric(),
                            errorrate=numeric())
student_list = unique(HCI1$Anon.Student.Id)
for(s in student_list){
  data = filter(HCI1,Anon.Student.Id==s)
  e1 <- data %>%
  select(Opportunity_Numeric, Success1) %>%
  group_by(Opportunity_Numeric) %>%
  summarise(error = 1- sum(Success1)/n())
  #e1=Error_calculate(data)
  student_error = rbind(student_error,data.frame(cbind(as.character(s),e1)),
                        use.names=FALSE)
}
```

```
data_pre = distinct(select(HCI1,c(Anon.Student.Id,TutorTime)))
data_pre =  data_pre[-c(4,6,12,26,30,40,44,45,52,59,60,72,78,93,103,106,109),]
#data_pre =  data_pre[-c(2,6,9,13,18,22,26,44,46,48,53,62,64,75,78,80,84,85,89,
#91,93,105,107,115,117,118,123,129,133,135,146,150,157,161,176,178,185,186,188,
```

```
#191,202,209),]
total <- merge(data_pre,student_error,by="Anon.Student.Id")
plot(density(data_pre$TutorTime),main = "Teacher Intervention Density Plot",
     xlab="Opportunity Number")
```




```
#plot(density(data_pre$TutorTime))
```

```
summary(data_pre$TutorTime)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   4.000   7.000   9.701  13.000  50.000
```

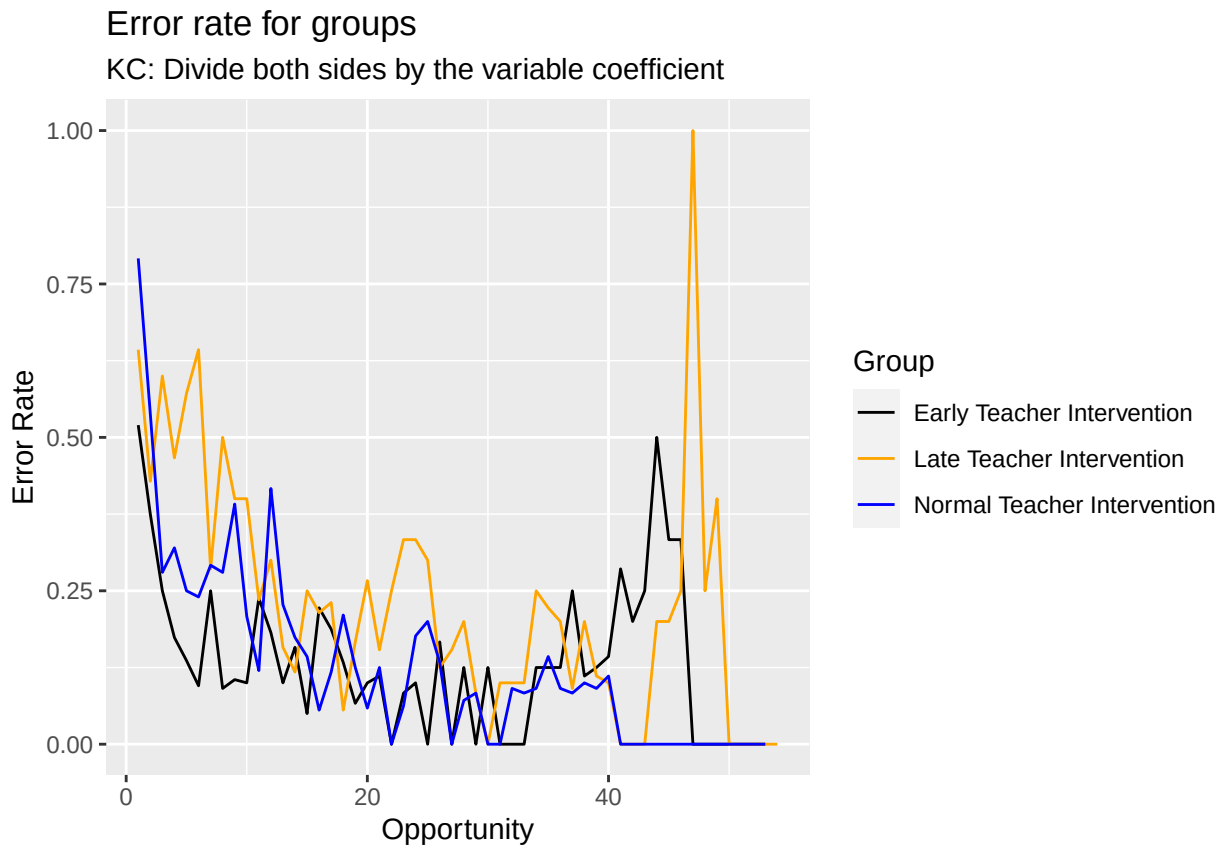
```
table(data_pre$TutorTime)
```

```
##
##  1  2  3  4  5  6  7  8  9 10 11 12 13 15 16 17 18 19 21 22 23 28 31 35 50
##  6  9  8  8  6  8  5  3  4  6  5  4  4  3  1  2  4  1  1  1  3  1  2  1  1
```

Base on the density plot, we choose tutor time <= 4 as early tutor group, 4 < tutor <= 11 as normal tutor group. tutor > 11 as late tutor group

```
library(nlme)
total = total%>%mutate(Group=ifelse(TutorTime<=4,"Early Teacher Intervention",
                                    ifelse(TutorTime>11,"Late Teacher Intervention",
                                           "Normal Teacher Intervention")))

#early:
early = filter(total,Group=="Early Teacher Intervention")
op_list = unique(early$opportunity)
early_table = data.table(opportunity=numeric(), errorrate=numeric())
for(o in op_list){
  o_data = filter(early,opportunity==o)
  o_er = mean(o_data$errorrate)
  early_table = rbind(early_table,data.frame(o,o_er),use.names=FALSE)
}

early_table$Group = "Early Teacher Intervention"
#late:
late = filter(total,Group=="Late Teacher Intervention")
op_list = unique(late$opportunity)
late_table = data.table(opportunity=numeric(), errorrate=numeric())
for(o in op_list){
  o_data = filter(late,opportunity==o)
  o_er = mean(o_data$errorrate)
  late_table = rbind(late_table,data.frame(o,o_er),use.names=FALSE)
}
late_table$Group = "Late Teacher Intervention"
#normal:
normal = filter(total,Group=="Normal Teacher Intervention")
op_list = unique(normal$opportunity)
normal_table = data.table(opportunity=numeric(), errorrate=numeric())
for(o in op_list){
  o_data = filter(normal,opportunity==o)
  o_er = mean(o_data$errorrate)
  normal_table = rbind(normal_table,data.frame(o,o_er),use.names=FALSE)
}
normal_table$Group = "Normal Teacher Intervention"

plot_df=rbind(early_table,normal_table,late_table)
```

```
ggplot(plot_df,aes(x=opportunity,y=errorrate,color=Group))+
  geom_line()+
  scale_colour_manual(values=c("black", "orange", "blue"))+
  labs(title="Error rate for groups",
       subtitle ="KC: Divide both sides by the variable coefficient" )+
  ylab("Error Rate")+
  xlab("Opportunity")
```



We see that the students with early tutor intervention have lower error rate than students with normal tutor intervention

New AFM for group

```
HCI11= HCI1
HCI11$Group = NA
for(s in unique(total$Anon.Student.Id)){
  list=which(HCI1$Anon.Student.Id %in% s)
  group_name = match(s,total$Anon.Student.Id)
  for(l in list){
    HCI11$Group[l] = total$Group[group_name]
  }
}
HCI11 <- HCI11[HCI11$Opportunity_Numeric<=40,]
```

```
L1 = length(HCI2$Anon.Student.Id)
Success1 = vector(mode="numeric", length=L1)
Success1[HCI11$First.Attempt=="correct"]=1
AFM11 = lmer(Success1~(1|Anon.Student.Id) + Opportunity_Numeric +
               Group:Opportunity_Numeric + Group,data=HCI11)
summary(AFM11)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula:
## Success1 ~ (1 | Anon.Student.Id) + Opportunity_Numeric + Group:Opportunity_Numeric +
```

```
##     Group
##    Data: HCI11
##
## REML criterion at convergence: 1300.8
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -3.1515 -0.1749  0.1525  0.4454  2.5908
##
## Random effects:
##  Groups          Name        Variance Std.Dev.
##  Anon.Student.Id (Intercept) 0.04989  0.2234
##  Residual                    0.10627  0.3260
## Number of obs: 1790, groups:  Anon.Student.Id, 97
##
## Fixed effects:
##                                                          Estimate Std. Error
## (Intercept)                                             7.760e-01  4.848e-02
## Opportunity_Numeric                                     6.011e-03  1.427e-03
## GroupLate Teacher Intervention                         -1.075e-01  7.265e-02
## GroupNormal Teacher Intervention                       -1.413e-01  6.643e-02
## Opportunity_Numeric:GroupLate Teacher Intervention     -5.781e-04  2.124e-03
## Opportunity_Numeric:GroupNormal Teacher Intervention -5.955e-06  1.897e-03
##                                                        t value
## (Intercept)                                             16.008
## Opportunity_Numeric                                      4.213
## GroupLate Teacher Intervention                          -1.480
## GroupNormal Teacher Intervention                        -2.127
## Opportunity_Numeric:GroupLate Teacher Intervention      -0.272
## Opportunity_Numeric:GroupNormal Teacher Intervention    -0.003
##
## Correlation of Fixed Effects:
##             (Intr) Oppr_N GrpLTI GrpNTI O_N:GLTI
## Opprtnty_Nm -0.384
## GrpLtTchrIn -0.667  0.256
## GrpNrmlTchI -0.730  0.280  0.487
## Oppr_N:GLTI  0.258 -0.672 -0.465 -0.188
## Oppr_N:GNTI  0.289 -0.752 -0.193 -0.395  0.505
```

We see that early intervention group has the highest success rate, followed by normal and late

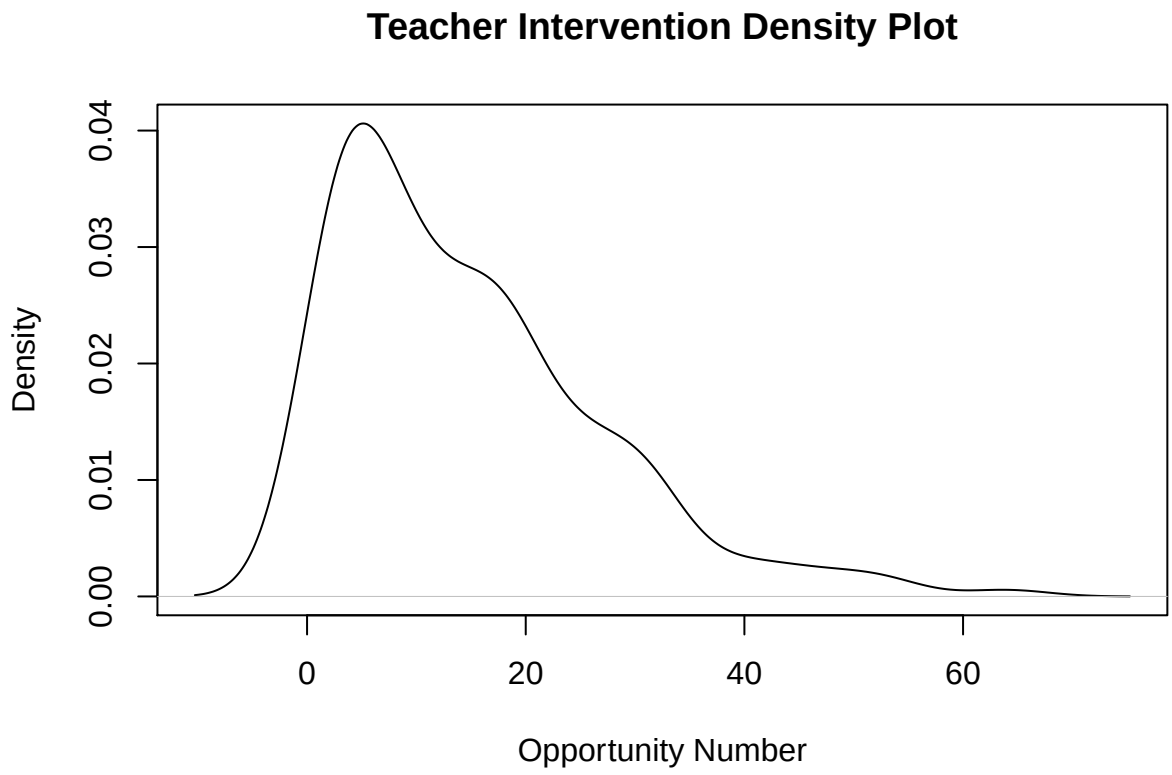
KC:Add constant terms

Divide students into early-tutor, normal-tutor, late-tutor groups.

```
library(data.table)
student_error <- data.table(Anon.Student.Id=character(),opportunity=numeric(),
                            errorrate=numeric())
student_list = unique(HCI2$Anon.Student.Id)
for(s in student_list){
  data = filter(HCI2,Anon.Student.Id==s)
  e1 <- data %>%
  select(Opportunity_Numeric, Success1) %>%
```

```
  group_by(Opportunity_Numeric) %>%
  summarise(error = 1- sum(Success1)/n())
  #e1=Error_calculate(data)
  student_error = rbind(student_error,data.frame(cbind(as.character(s),e1)),
                        use.names=FALSE)
}
```

```
data_pre = distinct(select(HCI2,c(Anon.Student.Id,TutorTime)))
data_pre =  data_pre[-c(4,6,12,26,30,40,44,45,52,59,60,72,78,93,103,106,109),]
#data_pre =  data_pre[-c(2,6,9,13,18,22,26,44,46,48,53,62,64,75,78,80,84,85,
#89,91,93,105,107,115,117,118,123,129,133,135,146,150,157,161,176,178,185,186,
#188,191,202,209),]
total <- merge(data_pre,student_error,by="Anon.Student.Id")
plot(density(data_pre$TutorTime),main = "Teacher Intervention Density Plot",
     xlab="Opportunity Number")
```




```
#plot(density(data_pre$TutorTime))
```

```
summary(data_pre$TutorTime)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1.00    5.00   12.00   14.73   21.00   64.00
```

```
table(data_pre$TutorTime)
```

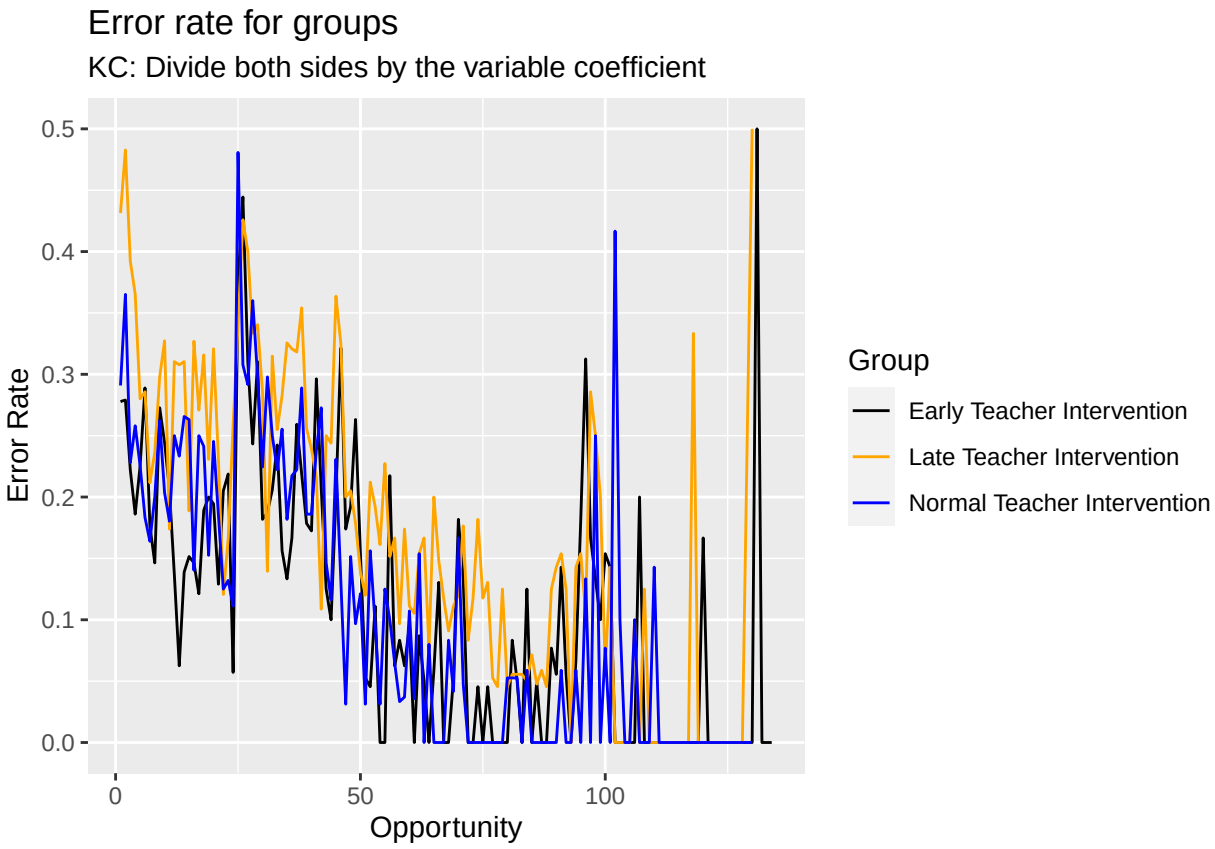
```
##
##  1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26
##  9 13 12 12  7 12  5  6  5  8  6  5  2  7  3  5  8 10  5  3  3  2  4  3  2  4
## 27 28 29 30 31 32 33 38 39 40 42 43 46 47 50 53 64
##  1  3  3  3  3  3  4  1  1  1  1  1  1  1  1  2  1
```

Base on the density plot, we choose tutor time <= 5 as early tutor group, 5 < tutor <= 17 as normal tutor group. tutor > 17 as late tutor group

```
library(nlme)
total = total%>%mutate(Group=ifelse(TutorTime<=5,"Early Teacher Intervention",
                                    ifelse(TutorTime>17,"Late Teacher Intervention",
                                           "Normal Teacher Intervention")))

#early:
early = filter(total,Group=="Early Teacher Intervention")
op_list = unique(early$opportunity)
early_table = data.table(opportunity=numeric(), errorrate=numeric())
for(o in op_list){
  o_data = filter(early,opportunity==o)
  o_er = mean(o_data$errorrate)
  early_table = rbind(early_table,data.frame(o,o_er),use.names=FALSE)
}

early_table$Group = "Early Teacher Intervention"
#late:
late = filter(total,Group=="Late Teacher Intervention")
op_list = unique(late$opportunity)
late_table = data.table(opportunity=numeric(), errorrate=numeric())
for(o in op_list){
  o_data = filter(late,opportunity==o)
  o_er = mean(o_data$errorrate)
  late_table = rbind(late_table,data.frame(o,o_er),use.names=FALSE)
}
late_table$Group = "Late Teacher Intervention"
#normal:
normal = filter(total,Group=="Normal Teacher Intervention")
op_list = unique(normal$opportunity)
normal_table = data.table(opportunity=numeric(), errorrate=numeric())
for(o in op_list){
  o_data = filter(normal,opportunity==o)
  o_er = mean(o_data$errorrate)
  normal_table = rbind(normal_table,data.frame(o,o_er),use.names=FALSE)
}
normal_table$Group = "Normal Teacher Intervention"

plot_df=rbind(early_table,normal_table,late_table)
```

```
ggplot(plot_df,aes(x=opportunity,y=errorrate,color=Group))+
  geom_line()+
```

```
  scale_colour_manual(values=c("black", "orange", "blue"))+
  labs(title="Error rate for groups"
       ,subtitle ="KC: Divide both sides by the variable coefficient" )+
  ylab("Error Rate")+
  xlab("Opportunity")
```



We see that the students with early tutor intervention have lower error rate than students with normal tutor intervention. This could be the reason why we observed the negative coefficients for the interaction term

New AFM for group

```
HCI22= HCI2
HCI22$Group = NA
for(s in unique(total$Anon.Student.Id)){
  list=which(HCI1$Anon.Student.Id %in% s)
  group_name = match(s,total$Anon.Student.Id)
  for(l in list){
    HCI22$Group[l] = total$Group[group_name]
  }
}
```

```
L1 = length(HCI22$Anon.Student.Id)
Success1 = vector(mode="numeric", length=L1)
Success1[HCI22$First.Attempt=="correct"]=1
AFM22 = lmer(Success1~(1|Anon.Student.Id) + Opportunity_Numeric +
```

```
              Group:Opportunity_Numeric + Group,data=HCI22)
summary(AFM22)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula:
## Success1 ~ (1 | Anon.Student.Id) + Opportunity_Numeric + Group:Opportunity_Numeric +
##     Group
##    Data: HCI22
##
## REML criterion at convergence: 1419.5
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -2.9262 -0.1020  0.2104  0.4848  2.1903
##
## Random effects:
##  Groups          Name        Variance Std.Dev.
##  Anon.Student.Id (Intercept) 0.04989  0.2234
##  Residual                    0.12122  0.3482
## Number of obs: 1735, groups:  Anon.Student.Id, 40
##
## Fixed effects:
##                                                    Estimate Std. Error
## (Intercept)                                       0.7268541  0.0499694
## Opportunity_Numeric                               0.0006762  0.0006844
## GroupLate Teacher Intervention                   -0.0404663  0.0447029
## GroupNormal Teacher Intervention                 -0.0566952  0.0473628
## Opportunity_Numeric:GroupLate Teacher Intervention  0.0001702  0.0008502
## Opportunity_Numeric:GroupNormal Teacher Intervention  0.0008143  0.0008681
##                                                   t value
## (Intercept)                                        14.546
## Opportunity_Numeric                                 0.988
## GroupLate Teacher Intervention                     -0.905
## GroupNormal Teacher Intervention                   -1.197
## Opportunity_Numeric:GroupLate Teacher Intervention  0.200
## Opportunity_Numeric:GroupNormal Teacher Intervention  0.938
##
## Correlation of Fixed Effects:
##             (Intr) Oppr_N GrpLTI GrpNTI O_N:GLTI
## Opprtnty_Nm -0.552
## GrpLtTchrIn -0.559  0.647
## GrpNrmlTchI -0.553  0.577  0.561
## Oppr_N:GLTI  0.452 -0.803 -0.804 -0.433
## Oppr_N:GNTI  0.449 -0.781 -0.496 -0.757  0.642
```

Early group has the highest success rate. The observations are similar to last KC