Factors of Successful Online Learning

Zoomers

Lillian Kim, Jemima Park, Jessica Yu

May 6, 2020

Abstract

Due to a recent pandemic of COVID-19, understanding the viability of remote learning is now more important than ever. We wanted to know if the factors that contribute to academic achievement under a traditional education system were still applicable to remote learning and how the remote learning could become a satisfying alternative for parents in the absence of traditional education. Multiple logistic regression analyses were conducted using an automated variable selection technique. Results showed that both student engagement in learning and parental attention to a child's education were just as important for a child's success and parent satisfaction in online learning as in a traditional mode of learning. Limitations and future directions are discussed.

Contents

Background and Significance

In light of recent pandemic of COVID-19, online learning has surged in popularity and necessity. As students who are currently adapting to online learning, we would like to assess what factors contribute to academic achievement, which we are defining as the overall grade given for the class. There are many factors that affect one's experience in a classroom, so what is essential in maximizing one's performance? In the traditional classroom setting factors such as participation and seeking resources contribute positively to overall performance. Does that translate over to online learning? There have been several studies published in the past few years in search of the best structure of online courses that generate high student satisfaction. A study by Ke and Xie (2009) found that students have higher overall satisfaction with their courses that include large amounts of discussion and organized course materials (as cited in Kauffman (2015)). What many of these studies lack is scope: most analyze results from one age group in one specific course. What we hope to do in our analysis is confirm findings in past studies regarding the association between high engagement and success in online class, as well as dig further to see if that relationship exists even across different subjects, age group, and gender.

On a related note, we are also interested in exploring parent satisfaction with online learning; recent events have shifted our idea of what learning can look like. We not only want to explore what attributes make online learning successful, but also how it compares in quality to learning in the traditional classroom setting. While we don't have the data to make a definitive comparison between the two methods, we can use the measure of parent satisfaction to do some initial exploration into this idea. If there is a signficant amount of parent satisfaction with online learning will more parents consider enrolling in these types of programs? Parent satisfaction is not a measure typically explored in previous studies, and we hope to shed light on this topic.

As time progresses, the nature of education as we know it is bound to change. Research on the effectiveness of alternative methods of learning is crucial for our society to adapt during unprecedented times.

Methods

Data collection

Kalboard 360, a learning management system, examined students' e-learning experience using a learner activity tracker tool that monitors learning processes and learner's behavioral engagement during online learning (Amrieh, Hamtini, and Aljarah (2016); Amrieh, Hamtini, and Aljarah (2015); Aljarah (2016)). From each student, Kalboard 360 also collected demographic information such as gender and nationality, as well as academic background information such as educational level and subject they took online learning for.

Each of 480 observations represents an individual student. Educational levels of students range from kindergarten to highschool. Although most students are from countries in the Middle East and Northern Africa, there are a few students outside of the region. We can generalize the results to the students of all educational levels in the Middle East and Northern Africa, and possibly to other regions as well.

Variable creation

Preliminary data wrangling

Unnecessary variables were removed from the dataset. *section_id* denoting the classroom a student belonged to was removed because it was unclear how the classrooms were decided in the original dataset. Variable *grade_id* that denoted which grade a student attended was removed because they provided information redundant to other variables. Variable *nationality* and *placeof_birth* were removed because some levels contained too few observations to conduct a proper, reliable analysis.

```r
# remove unnecessary variables
online <-
  subset(online.og, select=-c(nationality, placeof_birth, section_id, grade_id, semester))
```

Many of the nominal variables in the dataset had descriptive labels for levels. For example, *student_absence_days* had two levels, "Under-7" for students with equal to or less than 7 days of absence and "Above-7" for students with more than 7 days of absence. Also, our primary response variable *class* originally had three levels–"L" for a grade below 69, "M" for a grade of 70-89, and "H" for a grade of 90-100–which was not feasible for logistic regression analysis requiring a binary response variable. Therefore, for our analysis, nominal variables were coded to have 0-1 binary levels where appropriate.

```r
# code variables into 0 and 1 binary where appropriate
online <- online %>%
  mutate(student_absence_days = case_when(student_absence_days == "Under-7" ~ 0,
                                          student_absence_days == "Above-7" ~ 1),
         parent_answering_survey = case_when(parent_answering_survey == "Yes" ~ 1,
                                             parent_answering_survey == "No" ~ 0),
         parent_school_satisfaction = case_when(parent_school_satisfaction == "Good" ~ 1,
                                                parent_school_satisfaction == "Bad" ~ 0))

online <- online %>%
  mutate(class = case_when(class == "L"~0,
                           class == "M" | class == "H" ~ 1))
```

Response variables

1. class A final grade of a student at the end of the semester. (nominal: "1" for a grade of 70-100, "0" for a grade of 0-69)
2. parent_school_satisfaction Whether a parent was satisfied with the school (nominal: "1" for being satisfied, "0" for not being satisfied)

Predictor variables

1. Demographic variables

- gender Self-reported biological sex (nominal: "M" for male, "F" for female)

2. Academic background variables

- stage_id The academic levels of schools a student attends (nominal: "lowerlevel", "MiddleSchool", "HighSchool")
- topic The course subject for which a student participated in online learning (nominal: "Arabic", "Biology", "Chemistry", "English", "French", "Geology", "History", "IT", "Math", "Quran", "Science", "Spanish")

3. Student engagement variables

- raised_hands The total number of times a student raised hand in class in a semester (quantitative: 0-100)
- announcements_view The total number of times the student checked the new announcements on the web page in a semester (quantitative: 0-100)
- discussion The total number of times the student participated in discussion groups in a semester (quantitative: 0-100)

- visited_resources The total number of times a student visited the course content web page in a semester (quantitative: 0-100)
- student_absence_days The total number of days when a student was absent from class in a semester (nominal: “0” for below 7 days, “1” more than 7 days)

4. Parent background variable

- relation A parent who answered the survey provided by the school (nominal: “Mum”, “Father”)

Analytic methods

To check if the models have predictive power on a different sample from the one it was developed, the original dataset will be split into a testing sample and a holdout sample. We will use multiple logistic regression to study the association between passing an online course and various student engagement measures as well as demographics. We will use another multiple logistic regression model to study the association between parent satisfaction and a combination of student characeristics and engagement measures. The effectiveness of the models will be evaluated using the likelihood ratio test, and VIFs will be calculated to detect any multicollinearity between the predictors.

Results

Descriptive analysis

```r
engagement <-rbind(favstats(~visited_resources,data=online),
                   favstats(~raised_hands, data=online),
                   favstats(~announcements_view, data=online),
                   favstats(~discussion, data=online))

rownames(engagement) <- c("Visited Resources","Raised Hands",
                          "Viewed Announcements", "Discussion")

engagement
```

	min	Q1	median	Q3	max	mean	sd	n	missing
Visited Resources	0	20.00	65	84	99	54.79792	33.08001	480	0
Raised Hands	0	15.75	50	75	100	46.77500	30.77922	480	0
Viewed Announcements	0	14.00	33	58	98	37.91875	26.61124	480	0
Discussion	1	20.00	39	70	99	43.28333	27.63773	480	0

First, we wanted to get a general idea of what how our predictors were distributed. 64% of students identified as male, 36% female. Over 50% of students were at the middle school level compared to only 7% that were in highschool, and the courses they took varied widely across 12 different subjects.Student engagement actions including the number of times they visited resources and participated in discussion had a mean at about 40-50 times with the maximum count being 100 (see table above). Over 70% of students passed their courses with a grade of 69 or higher on a 100-point scale.
For the data relevant to parents, 60% of parents were satisfied with the overall online education that their child was enrolled in. Over half of parents took the time to respond to the surveys sent by the school. Note that the parent satisfaction response does not come from the surveys. These variables are separate from each other.

Bivariate analyses

Parents who did respond to the surveys tended to report being satisfied with the online curriculm, whereas parents who did not respond tended to report not being satisfied with the online curriculm. In addition, approximately 70% of parents who were satisfied with the curriculm had children who had fewer than 7 days of absence. Parent response to survey and student absent days may be possible signficant predictors in parent satisfaction, and will be explored in our logistic regression model.

```
tally(~parent_school_satisfaction | parent_answering_survey,
      data=online, format="percent")
```

```
                          parent_answering_survey
parent_school_satisfaction        0        1
                         0 69.04762 15.92593
                         1 30.95238 84.07407
```

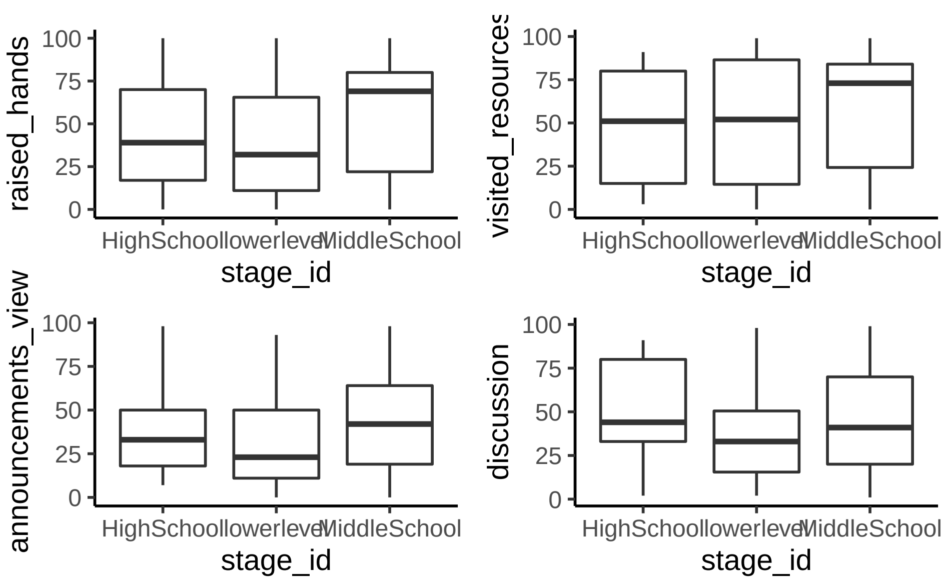
```
tally(~student_absence_days | parent_school_satisfaction,
      data=online, format="percent")
```

```
                    parent_school_satisfaction
student_absence_days        0        1
                   0 46.27660 69.17808
                   1 53.72340 30.82192
```

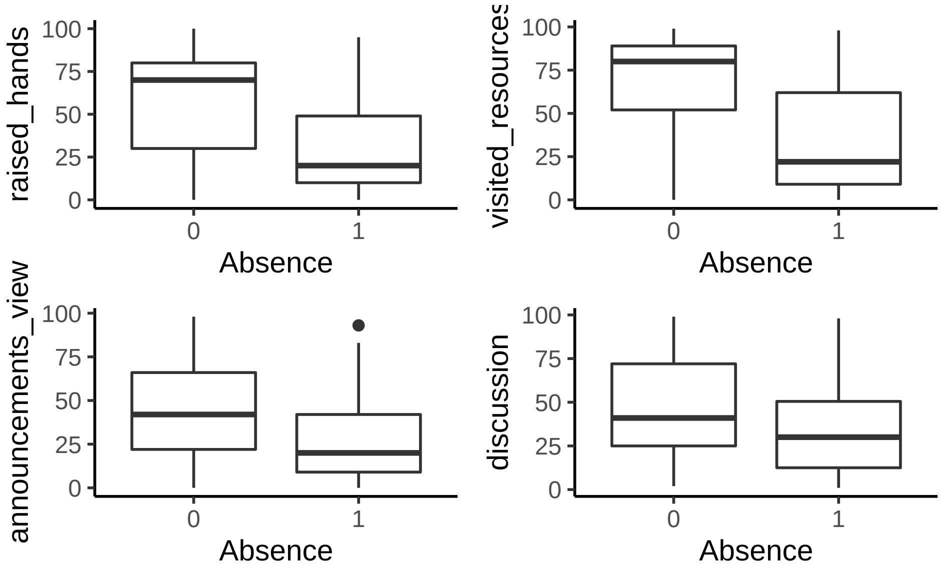
When the bivariate relationship includes at least one categorical predictor, we do not need to worry about VIFs or multicollinearity, but examining the relationship will help us gain some insight as to what kind of role these predictors might play in the logistic model.

Looking at grade level and various student engagement factors, the amount of student engagement does not vary across different ages. On the other hand, student engagement does seem to vary between students that were absent more frequently than not.

```
# student engagement by grade level
gridExtra::grid.arrange(r,v,a,d)
```




```
# student engagement by absence
gridExtra::grid.arrange(rr,vv,aa,dd)
```




```r
# Check multicollinearity issue among student engagement variables

## check correlations
check.corr <-
  subset(online, select = c("raised_hands", "discussion", "announcements_view",
                            "visited_resources"))

cor(check.corr)
```

```
                   raised_hands discussion announcements_view
raised_hands          1.0000000  0.3393860          0.6439178
discussion            0.3393860  1.0000000          0.4172900
announcements_view    0.6439178  0.4172900          1.0000000
visited_resources     0.6915717  0.2432918          0.5945000
                   visited_resources
raised_hands               0.6915717
discussion                 0.2432918
announcements_view         0.5945000
visited_resources          1.0000000
```

We suspected that the the number of times a student raised hand in a a semester, the number of times a sutdent participated in discussion groups in a semester, the number of times a student checked the new announcements on the web page in a semester, and the number of times a student visited the course content web page in a semester might be correlated to one another, because all four variables measured slightly different domains of student engagement in online learning. However, bivariate correlation analysis showed that there were no multicollinearity issues among these variables (all $rs < 0.70$). Therefore, all four variables were separately included in regression analysis without creating a composite variable of student engagement in learning.

Logistic regression analyses

```r
#randomly ordering observations
set.seed(3)
online2 <- online %>%
      mutate(random_num = rnorm(n = 480, mean=0, sd=1)) %>%
```

```
      arrange(random_num)

online2 <- subset(online2, select = -random_num)

#dividing the dataset into a training sample and holdout sample
online2.training <- online2[ c(1:240), ] #training
online2.holdout <- online2[ -c(1:240), ] #holdout
```

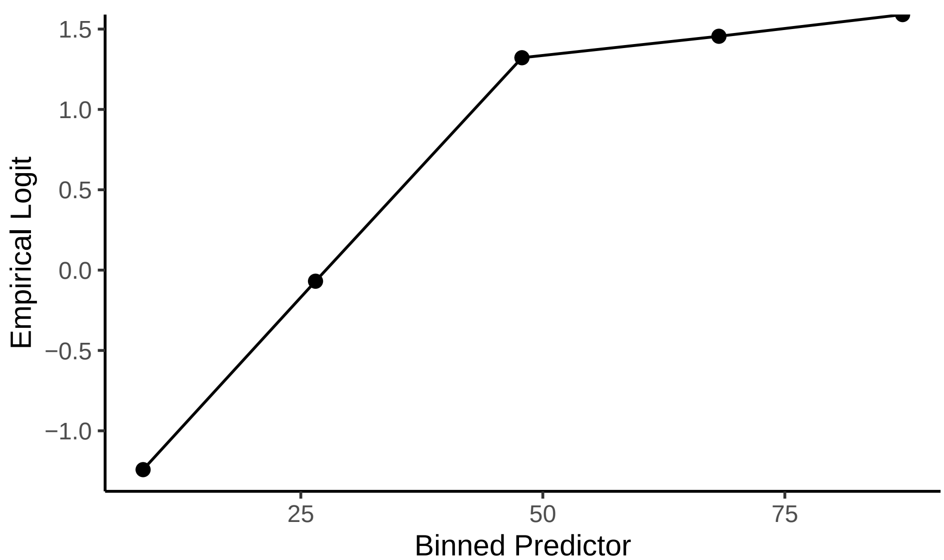
First, we randomly ordered 480 observations and separated the first 240 observations from the next 240 observations so that we could create a training sample and a holdout sample. We constructed a logistic regression model based on a training sample, and used the model on a holdout sample to test if the model had a predictive power for not only the sample it was based on but also other samples.

Inference Conditions

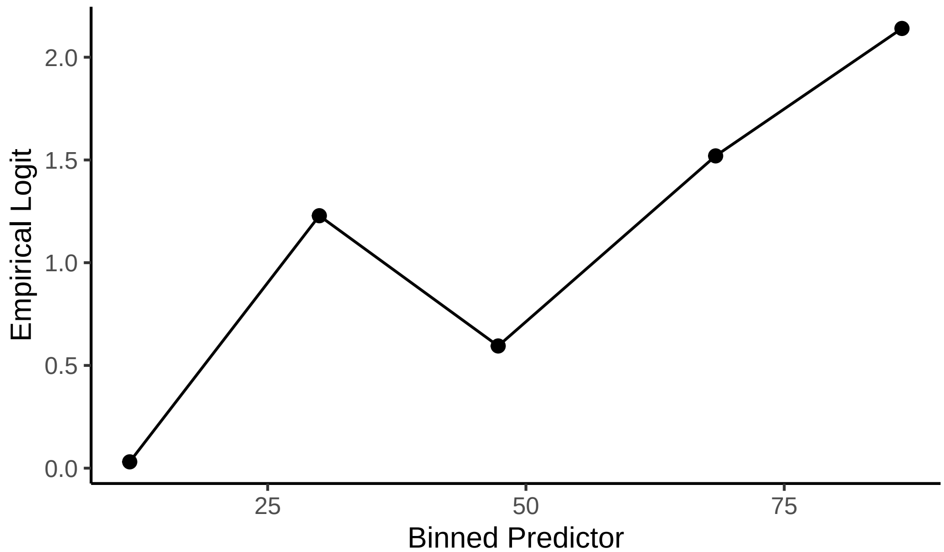
1. Linearity

We first checked the inference conditions for the model predicting whether the student achieves a passing grade with empirial logit plots.

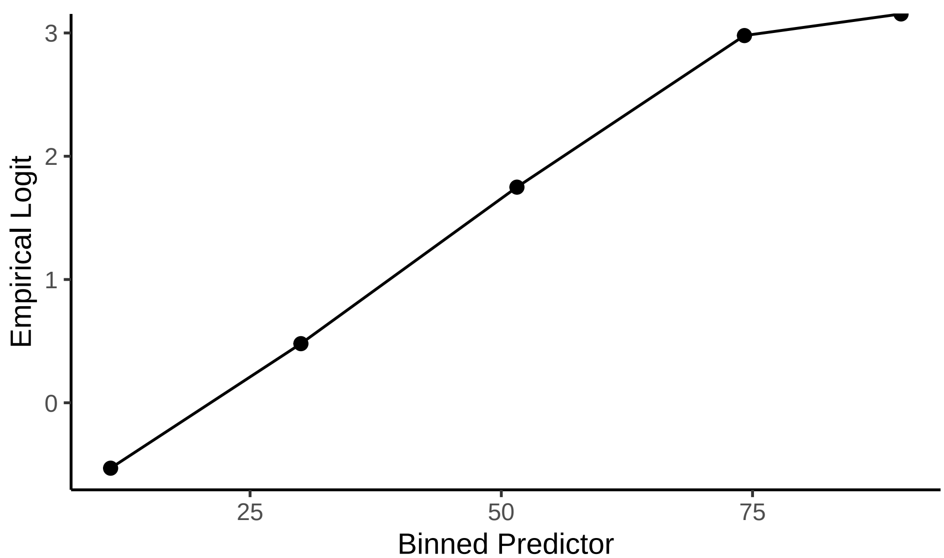
```
# the number of times a student visited resources web page
with(online2.training, emplogitplot(class, visited_resources, 5))
```



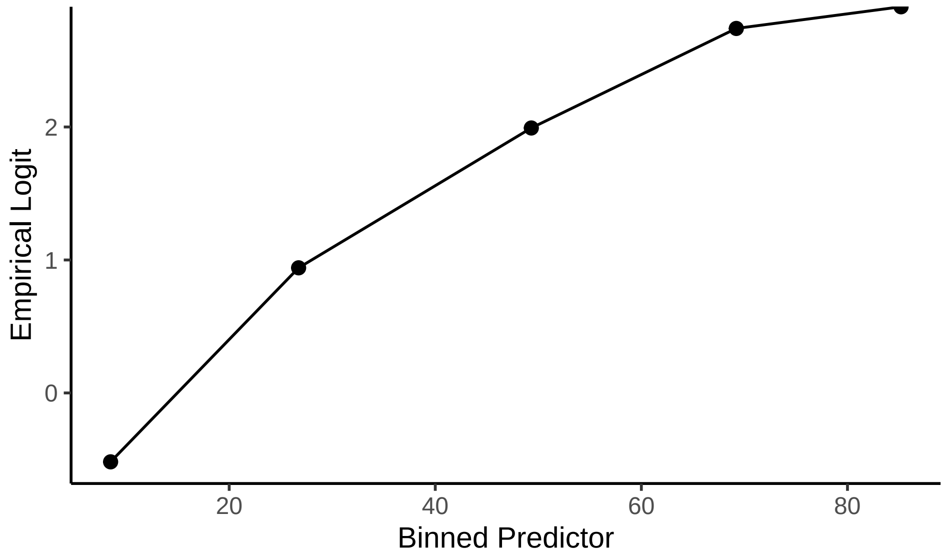

```
# the number of times a student participated in discussions
with(online2.training, emplogitplot(class, discussion, 5))
```




```r
# the number of times a student raised hand in class
with(online2.training, emplogitplot(class, raised_hands, 5))
```

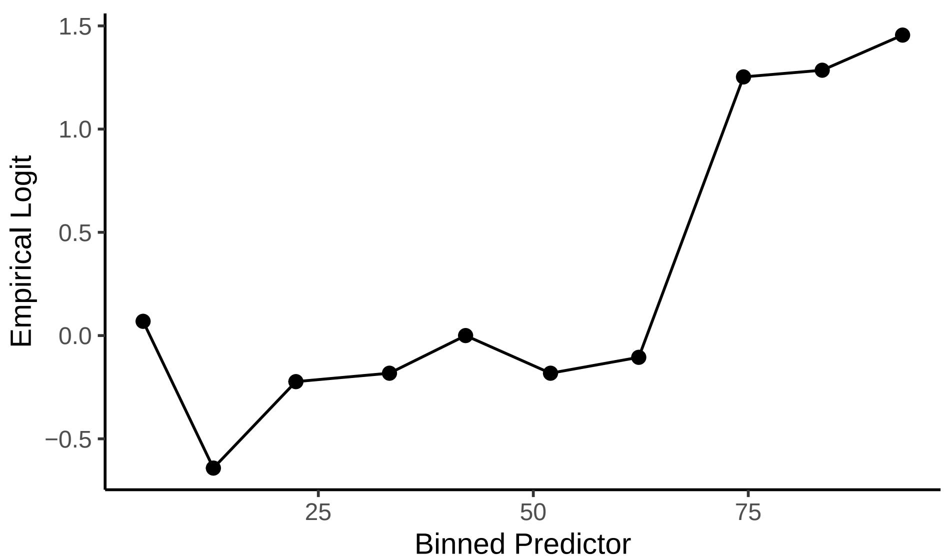



```r
# the number of times a student viewed new announcements
with(online2.training, emplogitplot(class, announcements_view, 5))
```

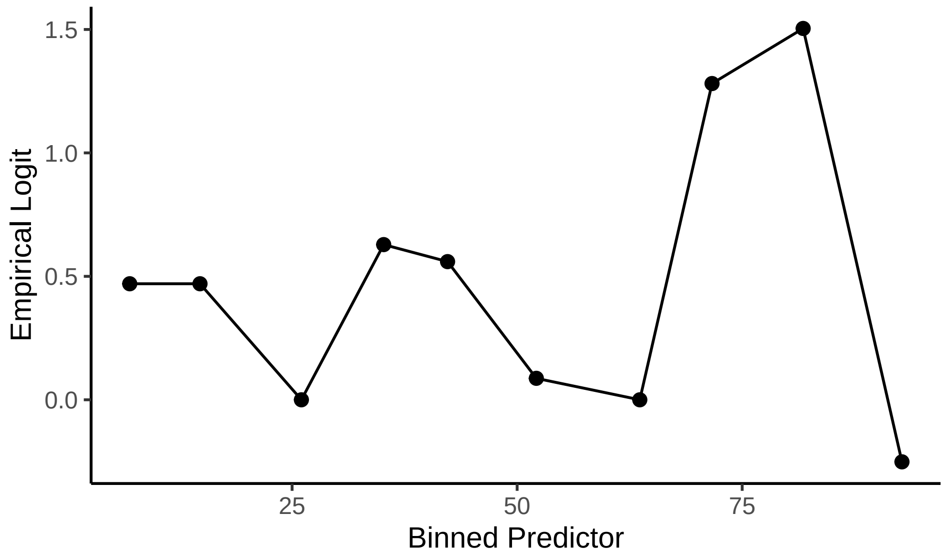


The linearity condition only needs to be checked for quantitative variables i.e. 4 variables measuring the student engagement. Based on the empirical logit plots, the number of times a student visited resources, the number of a student raised hand, and the number of times a student viewed announcements had a relatively linear relationship with log odds of a student passing a course. However, the number of times a student participated in discussions appeared to have a cubic relationship with the response variable. Therefore, we tried including a cubic term of *discussion* in the model predicting a pass/fail for a student below.

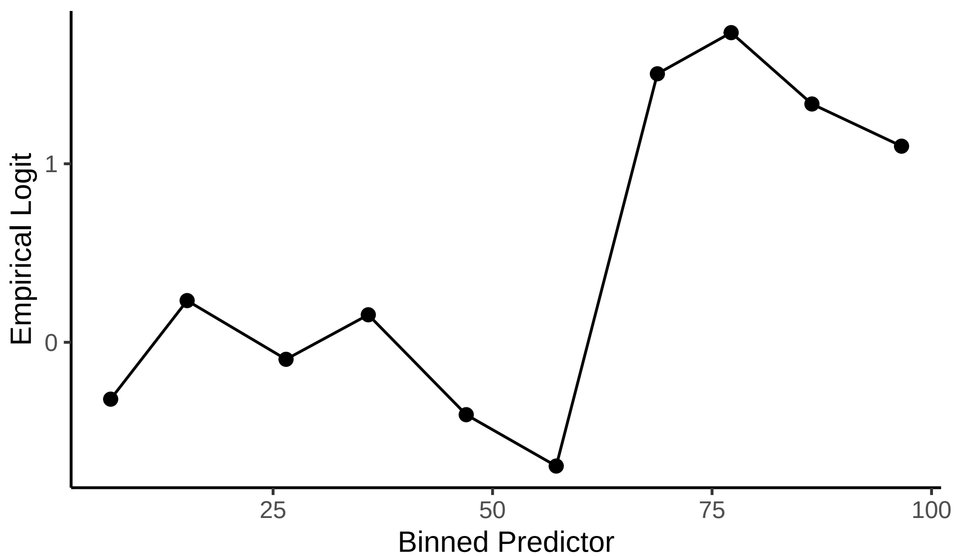
```
#For parent_school_satisfaction

# the number of times a student visited resources web page
with(online2.training, emplogitplot(parent_school_satisfaction, visited_resources, 10))
```



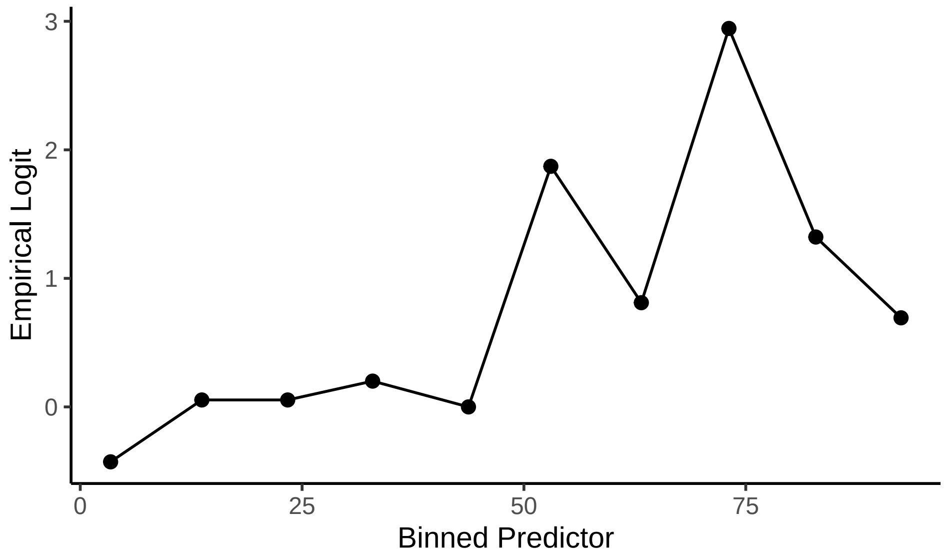

```
# the number of times a student participated in discussions
with(online2.training, emplogitplot(parent_school_satisfaction, discussion, 10))
```




```r
# the number of times a student raised hand in class
with(online2.training, emplogitplot(parent_school_satisfaction, raised_hands, 10))
```




```r
# the number of times a student viewed new announcements
with(online2.training, emplogitplot(parent_school_satisfaction, announcements_view, 10))
```



The empirical logit plots for whether a parent was satisfied with online schooling appeared far from linear, which suggested that there might not be a clear association between parental satisfaction and student engagement variables.

2. Independence and Randomness

The randomness condition for our logistic models are not met, as there was no mention of whether the data collected for student performance was a random sample. However, the independence condition is satisfied because student academic achivement is presumably independent from one another. While only a portion of our conditions were met, we proceeded with caution, making sure not to generalize our results to a wider population.

Automated variable selection

1. Predicting the pass/fail outcome of a student

Based on the empirical logit plot of *discussion*, which showed a cubic pattern, we included its squared and cubic terms in the automated variable selection process.

```
options(scipen= 999)
# Kitchen-sink model
k.sink.class <- glm(class~. + I(discussion^2) + I(discussion^3),
                    data=online2.training, family="binomial")

# Automated variable selection
auto.class <- stepAIC(k.sink.class, trace=FALSE)
msummary(auto.class)
```

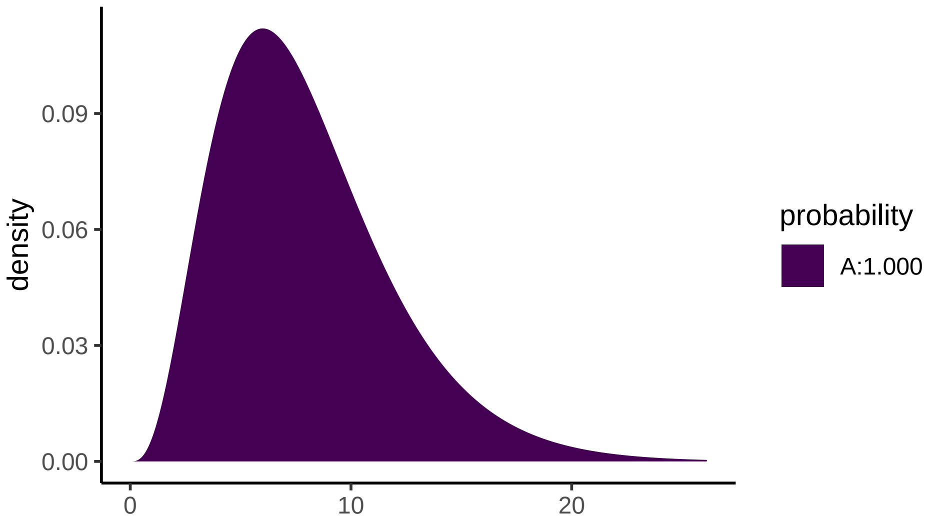
```
Coefficients:
                          Estimate Std. Error z value   Pr(>|z|)
(Intercept)             -0.5059223  1.0017504  -0.505    0.61353
genderM                 -1.0603621  0.7074673  -1.499    0.13392
raised_hands             0.0468334  0.0168399   2.781    0.00542 **
visited_resources        0.0250611  0.0119669   2.094    0.03624 *
announcements_view       0.0242042  0.0162572   1.489    0.13653
parent_answering_survey  1.9518353  0.6455515   3.024    0.00250 **
student_absence_days    -3.5827910  0.6645650  -5.391 0.00000007 ***
```

```
I(discussion^2)          0.0001655  0.0001162   1.424    0.15440

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 291.500  on 239  degrees of freedom
Residual deviance:  83.825  on 232  degrees of freedom
AIC: 99.825

Number of Fisher Scoring iterations: 8
```

```
#Include a linear term of discussion with a squared term
final.class <- glm(class ~ gender + raised_hands + visited_resources +
                     announcements_view + parent_answering_survey +
                     student_absence_days + discussion + I(discussion^2),
                   data=online2.training, family="binomial")
msummary(final.class)
```

```
Coefficients:
                           Estimate  Std. Error z value     Pr(>|z|)
(Intercept)             -0.64544062  1.20406891  -0.536      0.59192
genderM                 -1.07241045  0.70969185  -1.511      0.13076
raised_hands             0.04651483  0.01690248   2.752      0.00592 **
visited_resources        0.02530949  0.01203339   2.103      0.03544 *
announcements_view       0.02384765  0.01633025   1.460      0.14420
parent_answering_survey  1.92074328  0.66077415   2.907      0.00365 **
student_absence_days    -3.56718718  0.66786597  -5.341 0.0000000923 ***
discussion               0.00910014  0.04303349   0.211      0.83252
I(discussion^2)          0.00007684  0.00043501   0.177      0.85979

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 291.50  on 239  degrees of freedom
Residual deviance:  83.78  on 231  degrees of freedom
AIC: 101.78

Number of Fisher Scoring iterations: 8
```

```
#Likelihood Ratio Test (LRT) for model utility
xpchisq(final.class$null.deviance - final.class$deviance,
        df = final.class$df.null - final.class$df.residual)
```




```
[1] 1
```

```
#check multicollinearity
car::vif(final.class)
```

```
               gender             raised_hands        visited_resources 
             1.157783                 1.250944                 1.200023 
   announcements_view parent_answering_survey     student_absence_days 
             1.072678                 1.317712                 1.174170 
           discussion          I(discussion^2) 
            16.238083                16.993640 
```

Best subsets automated selection technique produced a model with *gender*, *raised_hands*, *visited_resources*, *discussion^2*, *announcements_view*, *student_absence_days*, and *parent_answering survey* as predictors. We constructed a final model by adding the linear term *discussion*, because it is necessary to include the linear term if its polynomial term is significant. In short, a student's gender, the degree to which a student was engaged in online learning, and whether a parent took the time to answer a survey from school together significantly predicted whether a student would pass the course at the end of the semester (G=207.59, df=8, p<.001, AIC=101.78). The VIFs of the predictors were all less than 5 except for the VIFs of the linear *discussion* term and the squared *discussion* term, which were expected to be highly correlated.

```
#interpret coefficients in odds ratio
exp(coefficients(final.class))
```

```
          (Intercept)                  genderM             raised_hands 
           0.52443142               0.34218271               1.04761361 
    visited_resources       announcements_view parent_answering_survey 
           1.02563249               1.02413428               6.82603021 
 student_absence_days               discussion          I(discussion^2) 
           0.02823516               1.00914168               1.00007685 
```

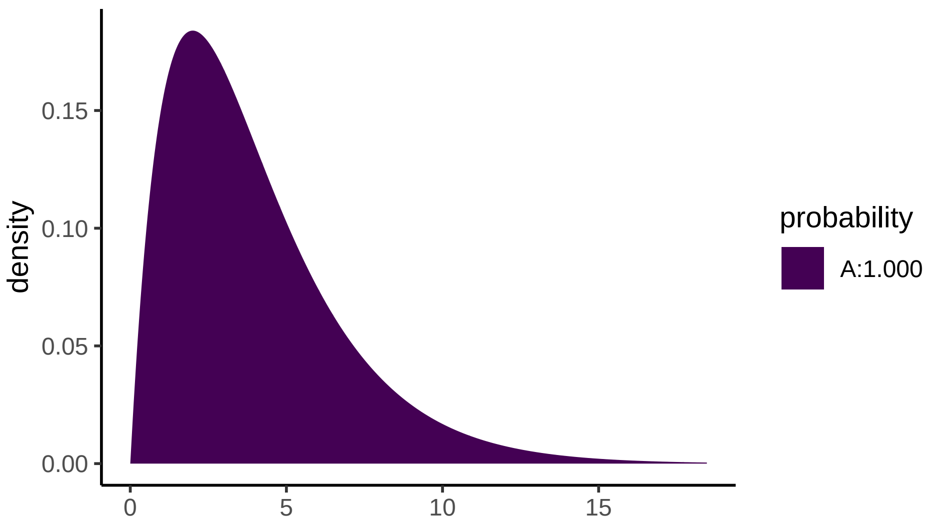
We can put some of these coefficents in context. Adjusting for all other characterisitcs, every additional time a student views the new announcement web page is associated with 2.4% increase in the odds of passing a class (OR=1.024). Every additional time a student raises hand in a class is associated with 4.7% increase in the odds of passing a class (OR=1.047). If a student is absent for more than 7 days in a course, the odds of passing a course decreases by 99.972% (OR=0.028). The odds of passing the course for a student whose parent answered a survey were 6.83 times the odds of passing a course for a student whose parent did not answer a survey (OR=6.826).

2. Predicting a parental satisfaction with online schooling

```
# Kitchen-sink model
k.sink.satis <- glm (parent_school_satisfaction ~ .,
                     data = online2.training, family = "binomial")

final.satis <- stepAIC(k.sink.satis, trace = FALSE)
msummary(final.satis)
```

```
Coefficients:
                         Estimate Std. Error z value          Pr(>|z|)
(Intercept)             -1.141866   0.356669  -3.201           0.00137 **
relationMum              0.851053   0.365149   2.331           0.01977 *
raised_hands             0.010965   0.006045   1.814           0.06970 .
discussion              -0.010081   0.006798  -1.483           0.13813
parent_answering_survey  2.535984   0.349484   7.256 0.000000000000398 ***

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 320.46  on 239  degrees of freedom
Residual deviance: 226.85  on 235  degrees of freedom
AIC: 236.85

Number of Fisher Scoring iterations: 4
```

```
#Likelihood Ratio Test (LRT) for model utility
xpchisq(final.satis$null.deviance-final.satis$deviance,
        df = final.satis$df.null - final.satis$df.residual)
```




```
[1] 1
```

```
#check multicollinearity
car::vif(final.satis)
```

```
               relation            raised_hands              discussion
               1.110700                1.261027                1.235538
parent_answering_survey
               1.091324
```

Best subsets automated selection technique produced a model with *relation*, *raised_hands*, *discussion*, and *parent_answering_survey* as significant predictors for whether a parent was satisfied with school. In short, which parent was primarily responsible for educational communication with school, whether a parent took time to answer a survey from school, and the number of times a student raised hands in class and participated in discussion together significantly predicted whether a parent was satisfied with the child's online schooling (G=93.61, df=4, p<.001, AIC=236.85). According to the VIFs of the predictors, there was no multicollinearity issue in this model.

```
exp(coefficients(final.satis))
```

```
            (Intercept)             relationMum            raised_hands 
              0.3192226               2.3421108               1.0110251 
             discussion parent_answering_survey 
              0.9899701              12.6288528 
```

We can put some of these coefficents in context. Adjusting for all other characterisitcs, the odds of a parent being satisfied with online schooling is 11.63 times higher for a parent who took the time to answer the survey from the school than for a parent who did not (OR=12.63). If mothers fill out the survey from the school, their odds of being satisfied with a child's online education are 2.34 times the odds if fathers answer the survey (OR=2.342).

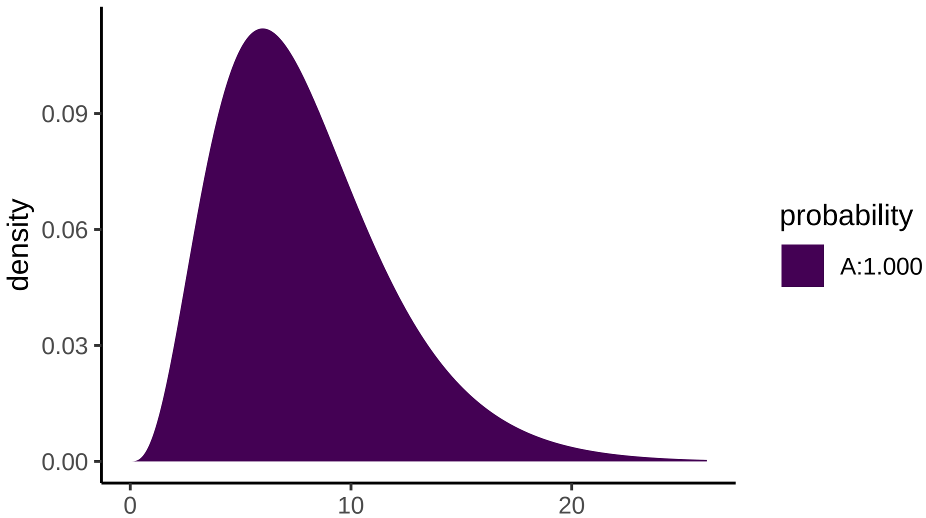
Testing the predictive power of the model

1. Predicting the pass/fail outcome of a student

```
# Test the model on a holdout sample
final.class.holdout <- glm(class ~ gender + raised_hands + visited_resources +
                     announcements_view + parent_answering_survey +
                     student_absence_days + discussion + I(discussion^2),
                   data=online2.holdout, family="binomial")

msummary(final.class.holdout)
```

```
Coefficients:
                          Estimate Std. Error z value   Pr(>|z|)    
(Intercept)             -2.2593269  1.1536965  -1.958    0.05019 .  
genderM                 -0.4158238  0.6410886  -0.649    0.51658    
raised_hands             0.0251861  0.0173247   1.454    0.14601    
visited_resources        0.0261706  0.0131846   1.985    0.04715 *  
announcements_view       0.0535318  0.0197642   2.709    0.00676 ** 
parent_answering_survey  1.0552850  0.6112893   1.726    0.08429 .  
student_absence_days    -2.9151284  0.6552487  -4.449 0.00000863 ***
discussion               0.1177074  0.0486276   2.421    0.01550 *  
I(discussion^2)         -0.0012926  0.0005028  -2.571    0.01015 *  

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 260.771  on 239  degrees of freedom
Residual deviance:  85.541  on 231  degrees of freedom
AIC: 103.54

Number of Fisher Scoring iterations: 7
```

```
xpchisq(final.class.holdout$null.deviance - final.class.holdout$deviance,
        df = final.class.holdout$df.null - final.class.holdout$df.residual)
```




```
[1] 1
```

Fitting the model on the holdout sample, we found that at least one of the predictors in the model was significant in predicting the pass/fail outcome of a student in the holdout sample (G=175.23, df=8, p<.001). Note that we cannot directly compare the deviance and AIC value because two models are based on different samples.

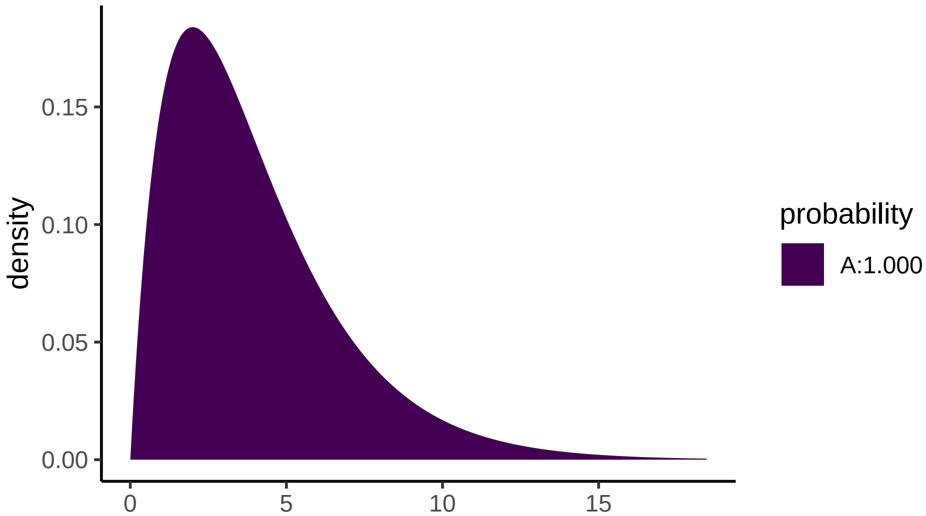
However, the changes in the significance of some predictors should be noted. Adjusting for all other predictors, the number of times a student raised hand and whether a parent answered a survey from school significantly predicted a pass/fail outcome in the training sample, but they were not significant anymore in the holdout sample (p=.146, p=.084). On the other hand, the number of times a student viewed announcements, which was not significant in the training sample, significantly predicted a pass/fail outcome in the holdout sample (p=.047). Both the linear and squared terms of the number of times a student participated in class discussions have also become significant predictors in the holdout sample, with p-values of p=.016 and p=0.010, respectively.

2. Predicting a parental satisfaction with online schooling

```
# Test the model on a holdout sample
final.satis.holdout <- glm(parent_school_satisfaction ~ relation + raised_hands +
                             discussion + parent_answering_survey,
                           data=online2.holdout, family="binomial")

msummary(final.satis.holdout)
```

```
Coefficients:
                         Estimate Std. Error z value         Pr(>|z|)
(Intercept)             -1.288009   0.353811  -3.640         0.000272 ***
relationMum              1.210073   0.361655   3.346         0.000820 ***
raised_hands             0.012588   0.006336   1.987         0.046930 *
discussion              -0.013705   0.007009  -1.955         0.050539 .
parent_answering_survey  2.475038   0.362238   6.833 0.00000000000834 ***

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 322.22  on 239  degrees of freedom
Residual deviance: 230.69  on 235  degrees of freedom
AIC: 240.69

Number of Fisher Scoring iterations: 4
```

```r
xpchisq(final.satis.holdout$null.deviance - final.satis.holdout$deviance,
        df = final.satis.holdout$df.null - final.satis.holdout$df.residual)
```




```
[1] 1
```

Testing the model on the holdout sample, we found that the model was still useful at predicting parental satisfaction with the same set of predictors (G=91.53, df=4, p<.001). In the training model, the student engagement variables such as the number of times a student raised hand and the number of times a student participated in discussions were only marginally significant (p=.070) or not significant (p=.138), respectively. But, the number of times a student raised hand was significant (p=.047), and the number of times a student participated in class discussions was marginally significant (p=.051) in predicting parental satisfaction in the holdout sample when all other predictors have been adjusted for in the model.

Conclusion

A recent pandemic of COVID-19 has necessitated remote learning for all levels of education, highlighting the importance of understanding how different the online learning is from the traditional method of learning. We wanted to know if the factors that contribute to academic achievement under a traditional education system were still applicable to remote learning. We also wanted to understand how to develop remote learning into a sustainable, viable alternative that could satisfy parents in the absence of traditional education.

Multiple logistic regression analyses were conducted to develop models predicting a pass/fail academic outcome of a student and parental satisfaction with online education. Just as in traditional classrooms, students who engaged more in the online learning by raising hand, checking daily announcements, utilizing resources, and participating in discussions were more likely to attain academic success. Especially, the number of days that a student was absent from online schooling was highly significant in predicting a student's pass/fail grade. A student with parents who were so attentive to a child's education as to complete a survey from school were more likely to succeed academically, which illustrated that a household environment and parental care was critical to academic success, regardless of the mode of learning.

We expected that a parent would be more satisfied with online learning if their child engaged more. Analysis revealed that student engagement was still relevant, albeit less, to predicting parental satisfaction. Mothers were more likely to report satisfaction with online schooling than fathers. Parents who care about their child's education enough to complete a survey from the school were more likely to be satisfied with online schooling. There are a few limitations in this analysis. First of all, the randomness condition is not fulfilled because the data was not collected through a random sampling. It is important to understand that the results cannot be generalized to a broader population, especially not to the college students and to the countries not in the

Middle East. Further research is necessary to generalize the findings to the American students doing the remote learning due to the current COVID-19 pandemic.

Another shortcoming of this dataset was that some variables which were likely measured quantitatively in the data collection process were already converted into categorical variables. For example, the number of days a student was absent from the online class in a semester was already converted into a categorical variable denoting whether a student was absent for more or less than 7 days. This way, a student who missed 8 days of an online school and a student who missed, say, 30 days of school are considered as equal in the analysis. More detailed and accurate analyses would be possible with raw data.

One direction a future research could pursue is to explore if there are any differences between different course subjects. Initially, we expected that the interaction of course subects and the level of school a student attends may be present because certain subjects, such as mathematics and chemistry, are known to become more difficult as students become older. However, such analysis was impossible because there were too few observations and too little variability in some cells. Future reserach should examine if online learning is more viable for certain course subjects than others by comparing average student academic achievement and parental satisfaction of, for example, STEM subjects and non-STEM subjects.

Overall, despite some limitations, analyses mostly confirmed our expectations that student engagement and parental attention to a child's education would be just as important in online learning as in a traditional mode of learning.

Appendix

1. Additional descriptive analyses

```r
# region
tally(~nationality, data=online.og, format = "percent")
```

```
nationality
      Egypt        Iran        Iraq      Jordan      Kuwait     Lebanon
  1.8750000   1.2500000   4.5833333  35.8333333  37.2916667   3.5416667
      Lybia     Morocco   Palestine SaudiArabia       Syria       Tunis
  1.2500000   0.8333333   5.8333333   2.2916667   1.4583333   2.5000000
        USA   Venezuela
  1.2500000   0.2083333
```

```r
# gender of participants
tally(~gender, data=online)
```

```
gender
  F   M
175 305
```

```r
tally(~gender, data=online, format="percent")
```

```
gender
       F        M
36.45833 63.54167
```

```r
# participant school level
tally(~stage_id, data=online, format="percent")
```

```
stage_id
  HighSchool   lowerlevel MiddleSchool
     6.87500     41.45833     51.66667
```

```r
# pass/fail outcome
tally(~class, data=online,format = "percent")
```

```
class
       0        1
26.45833 73.54167
```

```r
# course subjects
tally(~topic, data=online)
```

```
topic
   Arabic   Biology Chemistry   English    French   Geology   History
       59        30        24        45        65        24        19
       IT      Math     Quran   Science   Spanish
       95        21        22        51        25
```

```r
# Whether a parent was satisfied with online schooling
tally(~parent_school_satisfaction, data=online, format = "percent")
```

```
parent_school_satisfaction
       0        1
39.16667 60.83333
```

```r
# Whether a parent answered a survey from school
tally(~parent_answering_survey, data=online,format = "percent")
```

```
parent_answering_survey
    0     1
43.75 56.25
```

```r
# Which parent answered a survey from school
tally(~ relation , data=online,format = "percent")
```

```
relation
  Father      Mum
58.95833 41.04167
```

```r
# Whether a student missed more than 7 days of school
tally(~student_absence_days, data=online)
```

```
student_absence_days
  0   1
289 191
```

2. Interaction term

We expected that the interaction of whether a student was absent for more than 7 days and the level of school may be present because missing more days of school as a high school senior may impact success more than as a 1st grader. Therefore, we explored whether adding the `stage_id*student_absence_days` interaction term significantly improved either model.

```r
class.int1 <- glm(class ~ gender + raised_hands + visited_resources +
                    announcements_view + parent_answering_survey +
                    stage_id*student_absence_days + discussion + I(discussion^2),
                    data=online2.training, family="binomial")
msummary(class.int1)
```

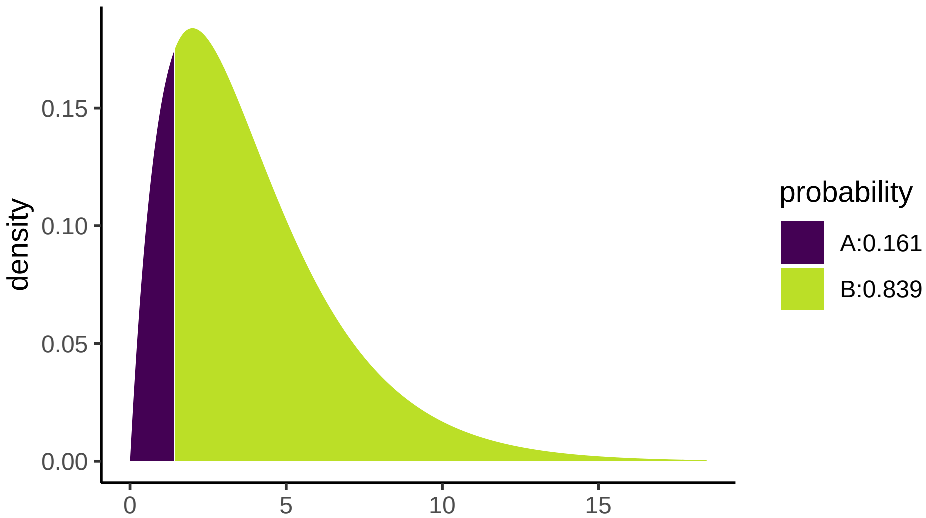
```
Coefficients:
                                              Estimate   Std. Error
```

```
(Intercept)                                  10.9545510 1287.1593745
genderM                                      -1.0292307    0.7128041
raised_hands                                  0.0483257    0.0174369
visited_resources                             0.0261481    0.0121072
announcements_view                            0.0261572    0.0169239
parent_answering_survey                       2.0965600    0.7020324
stage_idlowerlevel                          -11.5370601 1287.1587901
stage_idMiddleSchool                        -11.7807085 1287.1586826
student_absence_days                        -13.7429120 1287.1591027
discussion                                    0.0002822    0.0444930
I(discussion^2)                               0.0001902    0.0004550
stage_idlowerlevel:student_absence_days       9.7863036 1287.1597692
stage_idMiddleSchool:student_absence_days    10.1734016 1287.1595772

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 291.500  on 239  degrees of freedom
Residual deviance:  82.349  on 227  degrees of freedom
AIC: 108.35

Number of Fisher Scoring iterations: 16
```

```r
msummary(final.class)
```

```
Coefficients:
                           Estimate  Std. Error z value     Pr(>|z|)
(Intercept)             -0.64544062  1.20406891  -0.536      0.59192
genderM                 -1.07241045  0.70969185  -1.511      0.13076
raised_hands             0.04651483  0.01690248   2.752      0.00592 **
visited_resources        0.02530949  0.01203339   2.103      0.03544 *
announcements_view       0.02384765  0.01633025   1.460      0.14420
parent_answering_survey  1.92074328  0.66077415   2.907      0.00365 **
student_absence_days    -3.56718718  0.66786597  -5.341 0.0000000923 ***
discussion               0.00910014  0.04303349   0.211      0.83252
I(discussion^2)          0.00007684  0.00043501   0.177      0.85979

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 291.50  on 239  degrees of freedom
Residual deviance:  83.78  on 231  degrees of freedom
AIC: 101.78

Number of Fisher Scoring iterations: 8
```

```r
#nested LRT
xpchisq(final.class$deviance - class.int1$deviance, df=4)
```




```
[1] 0.1612668
```

Including the said interaction term did not significantly improve the prediction of a student's pass/fail outcome, p=.839.

```r
satis.int1 <- glm(parent_school_satisfaction ~ relation + raised_hands + discussion
                  + parent_answering_survey + stage_id*student_absence_days,
                    data= online2.training, family = "binomial")

msummary(satis.int1)
```

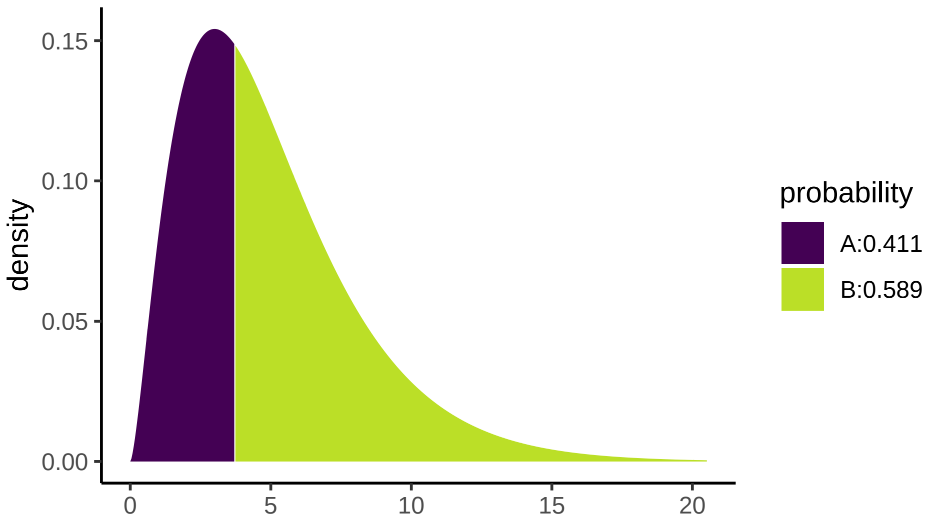
```
Coefficients:
                                           Estimate Std. Error z value
(Intercept)                                0.256645   1.225930   0.209
relationMum                                0.790486   0.370921   2.131
raised_hands                               0.009921   0.006812   1.456
discussion                                -0.011193   0.007081  -1.581
parent_answering_survey                    2.500400   0.359290   6.959
stage_idlowerlevel                        -0.820915   1.206475  -0.680
stage_idMiddleSchool                      -1.394116   1.176356  -1.185
student_absence_days                      -1.716108   1.625047  -1.056
stage_idlowerlevel:student_absence_days    0.996377   1.694886   0.588
stage_idMiddleSchool:student_absence_days  1.784174   1.672551   1.067

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 320.46  on 239  degrees of freedom
Residual deviance: 223.12  on 230  degrees of freedom
AIC: 243.12

Number of Fisher Scoring iterations: 5
```

```r
msummary(final.satis)
```

```
Coefficients:
                         Estimate Std. Error z value          Pr(>|z|)
(Intercept)             -1.141866   0.356669  -3.201           0.00137 **
relationMum              0.851053   0.365149   2.331           0.01977 *
raised_hands             0.010965   0.006045   1.814           0.06970 .
```

```
discussion               -0.010081   0.006798  -1.483            0.13813
parent_answering_survey  2.535984   0.349484   7.256 0.000000000000398 ***

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 320.46  on 239  degrees of freedom
Residual deviance: 226.85  on 235  degrees of freedom
AIC: 236.85

Number of Fisher Scoring iterations: 4
```

```r
#nested LRT
xpchisq(final.satis$deviance - satis.int1$deviance, df=5)
```




```
[1] 0.411187
```

Including the said interaction term did not significantly improve the prediction of a parental satisfaction with online learning, p=.589.

References

Aljarah, Ibrahim. 2016. "Students' Academic Performance Dataset." Kaggle. https://www.kaggle.com/aljarah/xAPI-Edu-Data.

Amrieh, E. A., T. Hamtini, and I. Aljarah. 2015. "Preprocessing and Analyzing Educational Data Set Using X-API for Improving Student's Performance." *Applied Electrical Engineering and Computing Technologies (AEECT)* 9 (8): 119–36. https://doi.org/10.1109/AEECT.2015.7360581.

———. 2016. "Mining Educational Data to Predict Student's Academic Performance Using Ensemble Methods." *International Journal of Database Theory and Application* 9 (8): 119–36. https://doi.org/10.14257/ijdta.2016.9.8.13.

Kauffman, H. 2015. "A Review of Predictive Factors of Student Success in and Satisfaction with Online Learning." *Research in Learning Technology* 23 (July): 1–13. https://doi.org/10.3402/rlt.v23.26507.