Tutoring Data Analysis

MAT 499

Dawson Damuth
Advised by Dr. Mark Baker

May 9, 2024

Introduction

The goal of my capstone research is to analyze a dataset of tutoring records to determine if there is significant evidence of an impact on an individual's grade in a course. Our question of interest is whether or not a final grade in a course is influenced by individuals being in a course section with an embedded tutor, as well as analyzing the influence of a number of variables.

Tutoring services at SUNY Oswego are offered through the Office of Learning Services (OLS). The services are provided at no cost and are available to the entire student body. The services offered are extensive and varied including one-on-one appointments, drop-in group tutoring, exam review sessions and embedded tutoring. The most recent addition to OLS' menu of services is Embedded Tutoring (ET). Embedded Tutoring provides tutorial assistance both in and out of the classroom creating a metaphorical bridge from the classroom to the tutoring center.

To assess this new service, OLS professional staff, together with the support of Institutional Research and Assessment have collected and compiled data concerning Embedded Tutor course status, tutoring visits, student demographic information and resulting course outcomes.

As an experienced peer tutor, Embedded Tutor and Applied Math Major, this research question was of particular interest to me. It is for this reason that I was asked to partner with OLS to work with fully anonymized data to construct and analyze a logistic regression model.

I would like to make note on behalf of Institutional Research and Assessment, the purpose of the project is to provide insight to the efficacy of the tutoring center and it's embedded tutor program, especially on bridging the achievement gap between under-represented minority (URM) and non-URM students, specifically in the field of mathematics. None of the information presented on the students is in any way identifiable, nor does it pose a risk to the individuals form which the data was gathered. The dataset was received with Student ID values randomized and demographic information included purely for the sake of our analysis.

Considering our data is from a unique collection of academic data, it is relevant to mention the difficulties academic research poses, as documented in a paper published by Roddy Theobald and Scott Freeman, "Is It the Intervention or the Students?". In their article, the two authors discuss the problems often associated with research in academia, most notably,

we expect the null hypothesis (no difference in our populations) to not be rejected in most cases. This is because, in these tests, there is no defined treatment category, as we cannot control which students belong to any given group when distinguished by other factors of interest, such as demographics. Working with tutoring data, even though all of the comparisons are made within a population of individuals that have attended tutoring, our sampling methods are not truly random, as we cannot control individuals who are and aren't offered tutoring. Further, when comparing among groups coming from classes with and without an embedded tutor, we cannot randomize who chooses what section to be in, and many people are not aware of what section has an embedded tutor. These issues cause our results to often be unreliable, and traditional methods of data analysis are not adequate for analyzing data sets of this nature. A second difficulty in analyzing the effectiveness of student interventions is the self-selected nature of participation. The ET model does not require students to participate with tutors and therefore attracts particular sub-groups of students in greater rates than others. These subgroups tend to be students who struggle in their course work, are retaking a course, are on academic probation, or lacking prerequisite skills for the course. Since it is natural for struggling students to seek tutoring; GPAs and course outcomes for tutored students are generally lower than those who do not attend tutoring. Teasing out the positive effect of tutoring attendance can be difficult.

Though I was unaware of the implications that using such a data set had when I began this project, my motivations were to investigate the effect the embedded tutoring program had on student course outcomes. Being an embedded tutor, myself, I found it intriguing to analyze a data set which allowed me to investigate the effects of a program which I have so much invested into. The hope was to find a significant result among the comparisons to conclude that the program is beneficial.

To demonstrate proficiency in the area of data analysis, I'd like to make mention of two of my previous projects which included analyzing Spotify records to predict whether a song will rank on the top charts using logistic regression, and a separate project using linear regression on a dataset of student demographic information to predict the student's final grade in a course. The relevant exposure in both projects has adequately prepared me to tackle this real-world dataset, selecting my own approach to do so, as well as demonstrate my understandings in the area of both statistical theory and applications.

Methodology

To investigate our questions of interest, we will implement a logistic regression model on the data, with the response being our passing criteria (C- or greater, P, S), and the predictors including embedded tutor status, tutoring duration in hours, and other corresponding student data. In place of linear regression, we use logistic regression in this scenario because of it's versatility in modeling data with a categorical response, allowing for categorical and numeric predictors, assessing the contribution and significance of each predictor in our model simultaneously. Logistic regression does have it's limitations in predictive power, most notably the requirement for randomly selected large samples. Despite an initial data set of 3041 observations, refining the set to contain only relevant inputs reduced the sample size a significant amount. This reduced sample size should be taken into consideration when analyzing results.

Though we do have other options available for the method of analysis, such as Chi-Square tests, the logistic regression model is able to account for as many variables as we wish to include, as well as continuous predictors, though the model may suffer as a result of over-fitting in some cases. Other methods like Chi-Square test of Independence or Partial Chi-Square approaches lack odds and the ability to analyze multiple variables at once, and is more often suited for two-way tables of strictly categorical variables.

For the sake of our analysis, our response variable will be binary, 1 indicating that the individual's grade in the course was within a passing range, 0 otherwise. It was determined to define grades of W and I as unsuccessful as F and U outcomes. Our model will take contributions by our predictor terms to the log-odds of our passing outcome. To make this result more digestible, I will be converting all log-odds to read as percentage change in standard odds by default using the conversion $exp(\beta_i) - 1$, where β_i is the coefficient on our predictor in the logistic regression model. The percentage change reads as:
There is a $(exp(\beta_i) - 1\%)$ increase (decrease, if negative) in the odds of an individual passing a course resulting from a one unit increase in x_i, all else held constant.

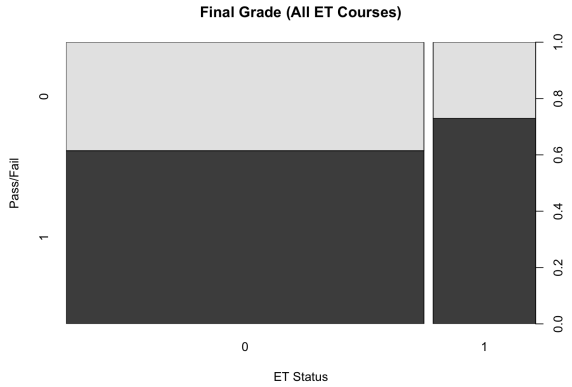
When optimizing our model performance, it is often of interest to create a threshold for our response as to determine when a response is considered success or fail (1 or 0). After building our models, we will use receiver operating characteristic (ROC) curves to determine this threshold, utilizing classification statistics to identify how effective the model is in predicting our data outcomes. This will be our final method of evaluating our model.

Before we begin our analysis, the dataset needs to be cleaned of bad observations and reduced to only include MAT courses. In addition, to meet the logistic regression criteria of independent observations, we must consolidate the data to have no repeated student identification numbers unless they are listed for attending tutoring for multiple courses. Consolidating the individual records meets the requirements of the logistic regression model and creates a total hours tutored variable. Unfortunately, this action also reduces the sample size to a final count of 431.

Results

Comparison Plots

The following plot compares ET status to course success in each category (Note: This comparison only includes courses which offer embedded tutors in one or more sections):



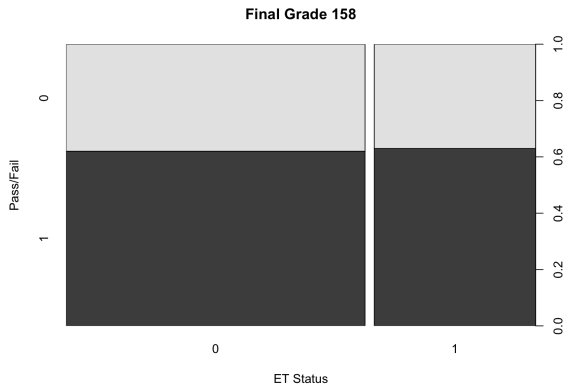
Reading the above graph, we can interpret that:

- The proportion of individuals in an embedded section is less than the proportion of those who are not.
- The proportion of individuals passing is $\sim 70\%$ among embedded students and $\sim 60\%$ among all other students.

Model Building: MAT158

Our first model will investigate MAT158 data, since it has the highest observation count, and is therefore most fit to apply logistic regression to. We will be using previous attempts in the course, embedded tutor status, and duration as our predictors.

MAT158 data produces the following individual comparison plot:



Here, we can see our plot suggests that ET presence doesn't make an impact on the performance of our sample.
Our main effects model for MAT158 is as follows:

$$logit[\hat{\pi}] = -0.035 - 0.169x_{et} + 0.138x_t + 0.254retakes$$

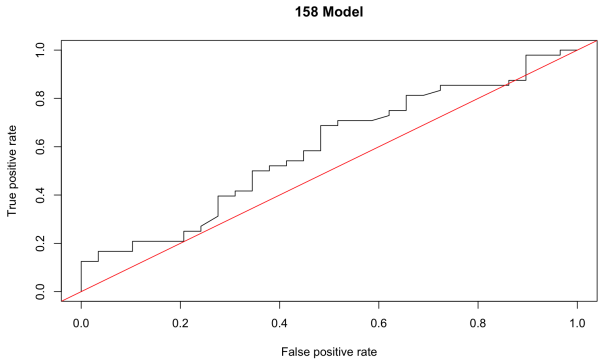
We can interpret the parameter estimates as:
The presence of an embedded tutor in a given course results in a $exp(-0.169) - 1 = -0.156$ 15.6% decrease in the odds of successful course completion, all else held constant.

A one hour increase in duration attending tutoring results in a 15.6% increase in the odds of successful course completion, all else held constant.

Every retake of a given course results in a 28.9% increase in the odds of successful course completion, all else held constant.

When performing a Likelihood Ratio test comparing our model with all predictors to the intercept model we get a test statistic value of $\chi^2 = Deviance_{full} - Deviance_{null} = 1.951$ on $df = df_{full} - df_{null} = 3$, which yields a p-value of 0.583. This result suggests that our model including predictors is no better at predicting success than a model which only includes the intercept term.

Next, our ROC curve for MAT158 is as follows:



Through software calculation, we can determine the optimal cutpoint for our response to be $\pi_0 = 0.59$. With this result, we can create our confusion matrix comparing our predictions to our data.

Prediction, $\pi_0 = 0.59$

Actual	$\hat{y} = 0$	$\hat{y} = 1$	
y=0	TN = 15	FP = 14	29
y=1	FN = 15	TP = 33	48
Total	30	47	77

This table allows us to compute the accuracy, sensitivity, and specificity. The results are as follows:

- Accuracy: (TP+TN)/(TP+FP+TN+FN)= 62.34%
- Sensitivity: TP/(TP+FN) = 68.75%
- Specificity: TN/(TN+FP) = 51.72%

We read these as; our model is 62% accurate in predicting our response, 69% accurate in predicting a student to have passed given they actually passed, and 52% accurate in predicting a student to have failed given they actually failed.

Validity of Results

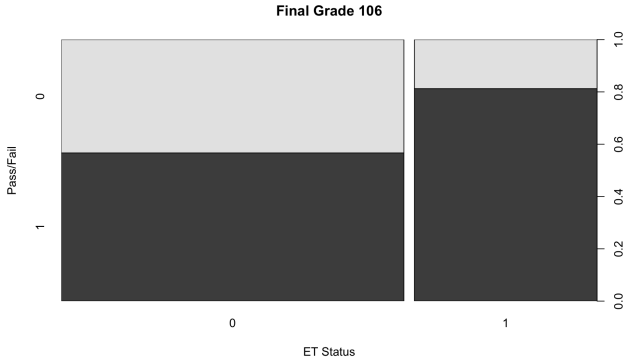
Though the above interpretations seem to indicate a strong influence on the response by our predictors, not a single predictor in our model has significance according to our p-values corresponding to the significance of our coefficients. Assumptions have not been met for logistic regression as well, potentially confounding our results.

Typically, the assumptions for logistic regression include measuring variance inflation factors (VIF's), a sufficiently large sample size. For our model, the VIF measures are standard (< 5), but the sample size is not sufficient, as our sample is expected to be at least $10 * 3/0.2 = 150$ observations, where our sample is 77.

Model Building: MAT106

Our second model will investigate MAT106 data, applying the same methods of model building analysis and evaluation.

MAT106 data produces the following individual comparison plot:



Here, we can see our plot suggests a stronger effect on our response due to ET presence, in comparison to our similar plot for MAT158.
Our main effects model for MAT106 is as follows:

$$logit[\hat{\pi}] = -0.447 + 0.772x_{et} + 0.146x_t + 0.480x_{retakes}$$

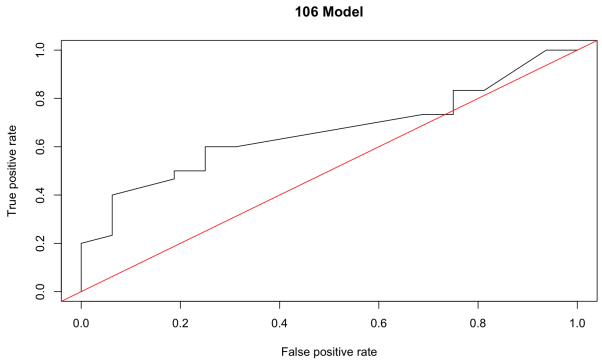
We can interpret the parameter estimates as:
The presence of an embedded tutor in a given course results in a 116.3% increase in the odds of successful course completion, all else held constant.

A one hour increase in duration attending tutoring results in a 15.7% increase in the odds of successful course completion, all else held constant.

Every retake of a given course results in a 61.6% increase in the odds of successful course completion, all else held constant.

When performing a Likelihood Ratio test comparing our model with all predictors to the intercept model we get a test statistic value of $\chi^2 = 4.827$ on $df = 3$, which yields a p-value of 0.185. This result suggests, again, that our model including predictors is no better at predicting success than a model which only includes the intercept term.

Next, our ROC curve for MAT106 is as follows:



Through calculation, we can determine the optimal cutpoint for our response to be $\pi_0 = 0.605$. With this result, we can create our confusion matrix comparing our predictions to our data.

Prediction, $\pi_0 = 0.605$

Actual	$\hat{y} = 0$	$\hat{y} = 1$	
y=0	TN = 12	FP = 4	16
y=1	FN = 12	TP = 18	30
Total	24	22	46

This table allows us to compute the accuracy, sensitivity, and specificity. The results are as follows:

- Accuracy: 65.22%
- Sensitivity: 60.00%
- Specificity: 75.00%

We read these as; our model is 65% accurate in predicting our response, 60% accurate in predicting a student to have passed given they actually passed, and 75% accurate in predicting a student to have failed given they actually failed.

Validity of Results

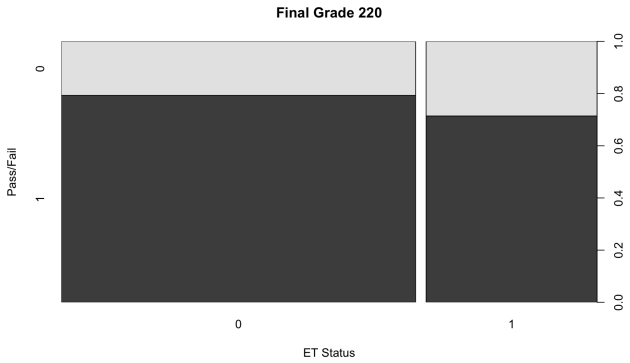
Similarly to our results from MAT158 data, not a single predictor in our model is significant according to our p-values corresponding to the significance of our coefficients. Assumptions have not been met for logistic regression as well, potentially confounding our results.

Measuring variance inflation factors (VIF's) for our model, the VIF measures are standard (< 5), but the sample size is not sufficient, as our sample is expected to be at least $10 * 3/0.2 = 150$ observations, where our sample is 46.

Model Building: MAT220

Our final model will investigate MAT220 data, applying the same methods of model building analysis and evaluation.

MAT220 data produces the following individual comparison plot:



Here, we can see our plot suggests an inversely strong effect on our response due to ET presence, in comparison to our similar plot for MAT106.

Our main effects model for MAT220 is as follows:

$$logit[\hat{\pi}] == 0.067 - 0.951x_{et} + 0.267x_t + 2.350x_{retakes}$$

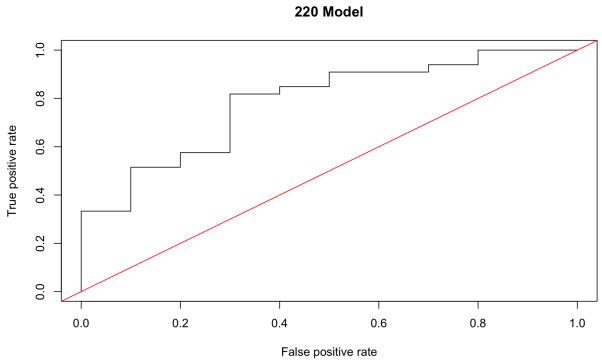
We can interpret the parameter estimates as:
The presence of an embedded tutor in a given course results in a 61.4% decrease in the odds of successful course completion, all else held constant.

A one hour increase in duration attending tutoring results in a 30.6% increase in the odds of successful course completion, all else held constant.

Every retake of a given course results in a 948.8% increase in the odds of successful course completion, all else held constant.

When performing a Likelihood Ratio test comparing our model with all predictors to the intercept model we get a test statistic value of $\chi^2 = 8.485$ on $df = 3$, which yields a p-value of 0.037. This result suggests that our model including predictors is finally better at predicting success than a model which only includes the intercept term.

Next, our ROC curve for MAT220 is as follows:



Through calculation, we can determine the optimal cutpoint for our response to be $\pi_0 = 0.64$. With this result, we can create our confusion matrix comparing our predictions to our data.

Prediction, $\pi_0 = 0.64$

Actual	$\hat{y} = 0$	$\hat{y} = 1$	
y=0	TN = 7	FP = 3	10
y=1	FN = 6	TP = 27	33
Total	13	30	43

This table allows us to compute the accuracy, sensitivity, and specificity. The results are as follows:

- Accuracy: 79.07%
- Sensitivity: 81.82%
- Specificity: 70.00%

We read these as; our model is 79% accurate in predicting our response, 82% accurate in predicting a student to have passed given they actually passed, and 70% accurate in predicting a student to have failed given they actually failed.

Validity of Results

Unlike our results from MAT158 and MAT106 data, our retakes predictor in our MAT220 model is significant according to our p-value (0.045). Though, assumptions have still not

been met for logistic regression, potentially confounding our results.

Measuring variance inflation factors (VIF's) for our model, the VIF measures are standard (< 5), but the sample size is not sufficient, as our sample is expected to be at least $10 * 3/0.2 = 150$ observations, where our sample is 43.

Conclusions

When considering the validity of our model, we generally want to ensure our model's predictive power is maximized given it meets assumption criteria, or alternatively, we can use cross-validation techniques to split our data into a training and testing set. Given the prior of these options isn't valid in any of our three models, we would normally consider the path of cross-validation. However, given our extremely small sample sizes, it makes it difficult to do so without sacrificing even more predictive power. This is one of the many complications our sample data presents. Similarly, when considering the type of data we have for our sample we also run into issues like those previously mentioned from "Is It the Intervention or the Students?". That being, when using institutional data we can never have a clear control group. This makes our data difficult to validate by assumption of randomness in our sample, and often contains very small sample sizes as a result of data relevance and availability in an institutional setting.

The complications we've mentioned so far are enough to diminish any hopes for model relevance. Though, it is my hope that these results may provide insight to the program's efficacy in some situations, regardless of overall significance in our results. Perhaps since we see that there is indication of an increase in grade given the number of retakes in all models, we can use this information to better advertise tutoring as a whole to those who are retaking courses for which it is offered, as well as embedded sections, specifically.

Returning to Theobald and Freeman's publication, there are a number of tests which may be better suited for this type of analysis. The publishers first mention a method of analyzing Raw Change Scores, which is simplified to performing a t-test to compare the means among our groups to determine if there is a significant difference between pre and post-scores. This method is contested by the publishers due to the inability to account for student's which may start with exceptionally low grades, and even improvements which may still be within failing ranges are considered greater improvement than a student in the control which may begin and end with a grade in the high-nineties. Other methods discussed include Normalized Gain Scores, Normalized Change Scores, and Effect Sizes, all of which are questioned in their applications for similar reasoning. Even so, these approaches appear to be better fit for our data than logistic regression, and could be considerations for our analysis, should it be taken any further.

References

Works Cited

Theobald, Roddy, and Scott Freeman. "Is It the Intervention or the Students? Using Linear Regression to Control for Student Characteristics in Undergraduate STEM Education Research." CBE-Life Sciences Education, 13 Oct. 2017, www.lifescied.org/doi/10.1187/cbe-13-07-0136.

Appendix

Variable Names, Descriptions and Types

The variables we will be using from the data set are:

- Student ID - Randomized student identification numbers, unique to each student.
- Duration (Hours) - Numeric beginning at zero representing the length of the individual appointments.
- Course - Numeric value associated with course catalog.
- ET - Indicator variable for whether the course section has an embedded tutor.
- Retakes - Number of previous attempts at the tutored course.
- Course Final Grade (Response) - Final grade in the tutored course, binary.

Variable Summaries

After our preprocessing is complete, we are able to continue to analyze our variables and obtain their 5-number summaries (for continuous variables) and frequency tables.

Course Counts

Course	Total Student Count	ET Section Count
101	20	6
102	61	9
104	44	8
106	46	16
120	48	8
158	77	27
179	10	0
206	2	0
208	30	8
210	39	0
215	3	0
220	43	14
230	2	0
240	4	0
249	2	0
258	19	0
318	1	0
330	1	0

Course Success

Success	freq.	rel. freq.
1	276	0.640
0	155	0.360

ET Section Status

ET	freq.	rel. freq.
1	96	0.223
0	335	0.777

Number of Retakes

Retakes	freq.	rel. freq.
0	183	0.425
1	213	0.494
2	32	0.074
3	3	0.007

R Code

```
#install.packages('dplyr')
#install.packages('psych')
#install.packages('DescTools')
#install.packages('tidyverse')
library(dplyr)
library(psych)
library(DescTools)
library(car)
library(ROCR)
library(caret)
library(tidyverse)
library(broom)

dat <- read.csv("MATTutoringData.csv", header=TRUE)

dat <- dat[dat$Subject == 'MAT',]

dat[is.na(dat)] <- 0

dat$retakes <- dat$Number.of.times.received.a.grade

dat$earned.1st <- dat$Earned.Grade.on.1st.attempt.at.Course

dat$earned.1st[dat$earned.1st == 'NO'] <- 0

dat$earned.1st[dat$earned.1st == 'YES'] <- 1


merged_data <- dat %>%
  group_by(Anonymous.ID, CRN, Course.Letter.Grade,
           Course.Numeric.Grade, Subject, Course,
           Ethnicity.Race, X1st.Gen, EOP, ED, URM,
           Gender,Accumulated.Hours.F23, ET.Section,
           Accumulated.Hours.S23, Credit.Load.S23,
           Credit.Load.F23, CGPA.S23, CGPA.F23, retakes,
           earned.1st) %>%
  summarise(hours = sum(Duration..Hours.))

summary(as.factor(merged_data$Course.Letter.Grade))

merged_data$Course.Numeric.Grade[merged_data$Course.Numeric.Grade >= 1.0] <- 1

merged_data$Course.Numeric.Grade[merged_data$Course.Numeric.Grade < 1.0] <- 0
```

```
freq.id <- table(as.factor(merged_data$Anonymous.ID))
summary(as.factor(freq.id))
summary(as.factor(freq.id))/sum(summary(as.factor(freq.id)))

hours.sum <- Desc(merged_data$hours, plotit = TRUE)
hours.sum

freq.crn <- table(as.factor(merged_data$CRN))
summary(as.factor(freq.crn))
summary(as.factor(freq.crn))/sum(summary(as.factor(freq.crn)))

freq.crn <- table(as.factor(merged_data$CRN))
as.factor(freq.crn)
freq.crn/sum(freq.crn)

freq.courseLG <- table(as.factor(merged_data$Course.Letter.Grade))
freq.courseLG
freq.courseLG/sum(freq.courseLG)

freq.success <- table(as.factor(merged_data$Course.Numeric.Grade))
freq.success
freq.success/sum(freq.success)

freq.eth.race <- table(as.factor(merged_data$Ethnicity.Race))
freq.eth.race
freq.eth.race/sum(freq.eth.race)

freq.1stgen <- table(as.factor(merged_data$X1st.Gen))
freq.1stgen
freq.1stgen/sum(freq.1stgen)

freq.EOP <- table(as.factor(merged_data$EOP))
freq.EOP
freq.EOP/sum(freq.EOP)

freq.ED <- table(as.factor(merged_data$ED))
freq.ED
freq.ED/sum(freq.ED)

freq.URM <- table(as.factor(merged_data$URM))
freq.URM
freq.URM/sum(freq.URM)

freq.gender <- table(as.factor(merged_data$Gender))
```

```
freq.gender
freq.gender/sum(freq.gender)

freq.ETSec <- table(as.factor(merged_data$ET.Section))
freq.ETSec
freq.ETSec/sum(freq.ETSec)

freq.retakes <- table(as.factor(merged_data$retakes))
freq.retakes
freq.retakes/sum(freq.retakes)

freq.earned <- table(as.factor(merged_data$earned.1st))
freq.earned
freq.earned/sum(freq.earned)

AccHrS.sum <- Desc(merged_data$Accumulated.Hours.S23, plotit = FALSE)
AccHrS.sum

AccHrF.sum <- Desc(merged_data$Accumulated.Hours.F23, plotit = FALSE)
AccHrF.sum

CrLoadS.sum <- Desc(merged_data$Credit.Load.S23, plotit = FALSE)
CrLoadS.sum

CrLoadF.sum <- Desc(merged_data$Credit.Load.F23, plotit = FALSE)
CrLoadF.sum

CGPAS.sum <- Desc(merged_data$CGPA.S23, plotit = TRUE)
CGPAS.sum

CGPAF.sum <- Desc(merged_data$CGPA.F23, plotit = FALSE)
CGPAF.sum

#Desc statistics for duration, final grade, prior grade, subsetted by class
# and ET presence. Graphical rep of each.

summary(as.factor(merged_data$Course))

data_et <- merged_data[merged_data$ET.Section == 1,]

summary(as.factor(data_et$Course))

summary_et <- data_et %>% group_by(Course) %>%
  summarise(observations = length(hours),
    mean(hours), sd(hours), mean(CGPA.F23), sd(CGPA.F23),
```

```
            mean(CGPA.S23), sd(CGPA.S23))

course <- as.factor(data_et$Course)

duration <- data_et$hours
plot(duration, course)

final_spring <- data_et$CGPA.S23
plot(final_spring, course)

final_fall <- data_et$CGPA.F23
plot(final_fall, course)

data_n_et <- merged_data[merged_data$ET.Section == 0,]

summary(as.factor(data_n_et$Course))

summary_n_et <- data_n_et %>% group_by(Course) %>%
  summarise(observations = length(hours), mean(hours), sd(hours), mean(CGPA.F23),
    sd(CGPA.F23),
            mean(CGPA.S23), sd(CGPA.S23))

course <- as.factor(data_n_et$Course)

duration <- data_n_et$hours
plot(duration, course)

final_spring <- data_n_et$CGPA.S23
plot(final_spring, course)

final_fall <- data_n_et$CGPA.F23
plot(final_fall, course)

#Compare ET presence to both grade and cumm. duration, separating by class,
# in addition to all MAT courses

#All MAT courses:

merged_data <- merged_data[merged_data$Course != 318
                           & merged_data$Course != 330
                           & merged_data$Course != 258,]

et_status <- as.factor(merged_data$ET.Section)

#numeric#
```

```
course_gpa <- as.factor(merged_data$Course.Numeric.Grade)

plot(y=course_gpa, x=et_status, main = 'Final Grade (All ET Courses)',
    ylab = 'Pass/Fail', xlab = 'ET Status')

duration <- merged_data$hours
plot(y=duration, x=et_status)

log.duration <- log(merged_data$hours)
plot(y=log.duration, x=et_status)

#By classes with ET presence:

summary(as.factor(data_et$Course))
#101
data.101 <- merged_data[merged_data$Course == 101,]

et_status <- as.factor(data.101$ET.Section)

course_gpa <- as.factor(data.101$Course.Numeric.Grade)

plot(y=course_gpa, x=et_status, main = 'Final Grade 101',
    ylab = 'Pass/Fail', xlab = 'ET Status')

duration <- data.101$hours
plot(y=duration, x=et_status)

log.duration <- log(data.101$hours)
plot(y=log.duration, x=et_status)

#102
data.102 <- merged_data[merged_data$Course == 102,]

et_status <- as.factor(data.102$ET.Section)

course_gpa <- as.factor(data.102$Course.Numeric.Grade)

plot(y=course_gpa, x=et_status, main = 'Final Grade 102',
    ylab = 'Pass/Fail', xlab = 'ET Status')

duration <- data.102$hours
plot(y=duration, x=et_status)

log.duration <- log(data.102$hours)
plot(y=log.duration, x=et_status)
```

```
#104
data.104 <- merged_data[merged_data$Course == 104,]

et_status <- as.factor(data.104$ET.Section)

course_gpa <- as.factor(data.104$Course.Numeric.Grade)

plot(y=course_gpa, x=et_status, main = 'Final Grade 104',
    ylab = 'Pass/Fail', xlab = 'ET Status')

duration <- data.104$hours
plot(y=duration, x=et_status)

log.duration <- log(data.104$hours)
plot(y=log.duration, x=et_status)

#106
data.106 <- merged_data[merged_data$Course == 106,]

et_status <- as.factor(data.106$ET.Section)

course_gpa <- as.factor(data.106$Course.Numeric.Grade)

plot(y=course_gpa, x=et_status, main = 'Final Grade 106',
    ylab = 'Pass/Fail', xlab = 'ET Status')

duration <- data.106$hours
plot(y=duration, x=et_status)

log.duration <- log(data.106$hours)
plot(y=log.duration, x=et_status)

#120
data.120 <- merged_data[merged_data$Course == 120,]

et_status <- as.factor(data.120$ET.Section)

course_gpa <- as.factor(data.120$Course.Numeric.Grade)

plot(y=course_gpa, x=et_status, main = 'Final Grade 120',
    ylab = 'Pass/Fail', xlab = 'ET Status')

duration <- data.120$hours
plot(y=duration, x=et_status)
```

```
log.duration <- log(data.120$hours)
plot(y=log.duration, x=et_status)

#158
data.158 <- merged_data[merged_data$Course == 158,]

et_status <- as.factor(data.158$ET.Section)

course_gpa <- as.factor(data.158$Course.Numeric.Grade)

plot(y=course_gpa, x=et_status, main = 'Final Grade 158',
    ylab = 'Pass/Fail', xlab = 'ET Status')

duration <- data.158$hours
plot(y=duration, x=et_status)

log.duration <- log(data.158$hours)
plot(y=log.duration, x=et_status)

#208
data.208 <- merged_data[merged_data$Course == 208,]

et_status <- as.factor(data.208$ET.Section)

course_gpa <- as.factor(data.208$Course.Numeric.Grade)

plot(y=course_gpa, x=et_status, main = 'Final Grade 208',
    ylab = 'Pass/Fail', xlab = 'ET Status')

duration <- data.208$hours
plot(y=duration, x=et_status)

log.duration <- log(data.208$hours)
plot(y=log.duration, x=et_status)

#220
data.220 <- merged_data[merged_data$Course == 220,]

et_status <- as.factor(data.220$ET.Section)

course_gpa <- as.factor(data.220$Course.Numeric.Grade)

plot(y=course_gpa, x=et_status, main = 'Final Grade 220',
    ylab = 'Pass/Fail', xlab = 'ET Status')
```

```
duration <- data.220$hours
plot(y=duration, x=et_status, main = 'Duration (Hours)',
    ylab = 'Duration', xlab = 'ET Status')

log.duration <- log(data.220$hours)
plot(y=log.duration, x=et_status, main = 'Log Duration (Hours)',
    ylab = 'Duration', xlab = 'ET Status')

summary(as.factor(data_et$Course))

summary(as.factor(merged_data$Course))

grade <- data.158$Course.Numeric.Grade

retakes <- data.158$retakes

et <- data.158$ET.Section

prior_gpa <- data.158$CGPA.S23

duration_ <- data.158$hours

mod <- glm(grade~et+duration_+retakes, data = data.158,
    family = binomial (link = "logit"))

(result <- summary(mod))

(exp(coef(mod))-1)*100

vif(mod)

png(filename = '158diagnostics.png')
par(mfrow=c(2,2))
plot(mod)
dev.off()

D0 <- result$null.deviance

D1 <- result$deviance

df.null <- result$df.null

df.full <- result$df.residual
```

```
1-pchisq(D0-D1,df=(df.null-df.full))

pred <- predict(mod, data.158, type='response')

bacc <- 0

for(k in 0:100){
  pred.class <- as.numeric(pred >= k/100)

  tp <- sum(pred.class == 1 & pred.class == data.158$Course.Numeric.Grade)

  fp <- sum(pred.class == 1 & pred.class != data.158$Course.Numeric.Grade)

  tn <- sum(pred.class == 0 & pred.class == data.158$Course.Numeric.Grade)

  fn <- sum(pred.class == 0 & pred.class != data.158$Course.Numeric.Grade)

  sens <- tp/(tp+fn)

  spec <- tn/(tn+fp)

  bacc[k] <- (sens+spec)/2
}

bacc

which(bacc==max(bacc))

pred.class <- as.numeric(pred >= which(bacc==max(bacc))/100)

confusionMatrix(factor(pred.class), factor(data.158$Course.Numeric.Grade),
    positive = '1')

roc.pred <- prediction(pred, data.158$Course.Numeric.Grade)

roc.perf <- performance(roc.pred, 'tpr', 'fpr')

plot(roc.perf, main='158 Model')
abline(0, 1, col='RED')

performance(roc.pred, 'auc')@y.values[[1]]

pred.test <- predict(mod, data.158, type = 'response')

pred.class.test <- as.numeric(pred.test >= which(bacc==max(bacc))/100)
```

```
accuracy.1 <- confusionMatrix(factor(pred.class.test),
    factor(data.158$Course.Numeric.Grade))$overall[1]

grade <- data.106$Course.Numeric.Grade

retakes <- data.106$retakes

et <- data.106$ET.Section

prior_gpa <- data.106$CGPA.S23

duration_ <- data.106$hours

mod <- glm(grade~et+duration_+retakes, data = data.106,
    family = binomial (link = "logit"))

(result <- summary(mod))

(exp(coef(mod))-1)*100

vif(mod)

png(filename = '106diagnostics.png')
par(mfrow=c(2,2))
plot(mod)
dev.off()

D0 <- result$null.deviance

D1 <- result$deviance

df.null <- result$df.null

df.full <- result$df.residual

1-pchisq(D0-D1,df=(df.null-df.full))

pred <- predict(mod, data.106, type='response')

bacc <- 0

for(k in 0:100){
  pred.class <- as.numeric(pred >= k/100)
```

```
  tp <- sum(pred.class == 1 & pred.class == data.106$Course.Numeric.Grade)

  fp <- sum(pred.class == 1 & pred.class != data.106$Course.Numeric.Grade)

  tn <- sum(pred.class == 0 & pred.class == data.106$Course.Numeric.Grade)

  fn <- sum(pred.class == 0 & pred.class != data.106$Course.Numeric.Grade)

  sens <- tp/(tp+fn)

  spec <- tn/(tn+fp)

  bacc[k] <- (sens+spec)/2
}

bacc

which(bacc==max(bacc))

pred.class <- as.numeric(pred >= which(bacc==max(bacc))/100)

confusionMatrix(factor(pred.class),
    factor(data.106$Course.Numeric.Grade), positive = '1')

roc.pred <- prediction(pred, data.106$Course.Numeric.Grade)

roc.perf <- performance(roc.pred, 'tpr', 'fpr')

plot(roc.perf, main='106 Model')
abline(0, 1, col='RED')

performance(roc.pred, 'auc')@y.values[[1]]

pred.test <- predict(mod, data.106, type = 'response')

pred.class.test <- as.numeric(pred.test >= which(bacc==max(bacc))/100)

accuracy.2 <- confusionMatrix(factor(pred.class.test),
    factor(data.106$Course.Numeric.Grade))$overall[1]

grade <- data.220$Course.Numeric.Grade

retakes <- data.220$retakes

et <- data.220$ET.Section
```

```
prior_gpa <- data.220$CGPA.S23

duration_ <- data.220$hours

mod <- glm(grade~et+duration_+retakes, data = data.220,
    family = binomial (link = "logit"))

(result <- summary(mod))

(exp(coef(mod))-1)*100

vif(mod)

png(filename = '220diagnostics.png')
par(mfrow=c(2,2))
plot(mod)
dev.off()

D0 <- result$null.deviance

D1 <- result$deviance

df.null <- result$df.null

df.full <- result$df.residual

1-pchisq(D0-D1,df=(df.null-df.full))

pred <- predict(mod, data.220, type='response')

bacc <- 0

for(k in 0:100){
  pred.class <- as.numeric(pred >= k/100)

  tp <- sum(pred.class == 1 & pred.class == data.220$Course.Numeric.Grade)

  fp <- sum(pred.class == 1 & pred.class != data.220$Course.Numeric.Grade)

  tn <- sum(pred.class == 0 & pred.class == data.220$Course.Numeric.Grade)

  fn <- sum(pred.class == 0 & pred.class != data.220$Course.Numeric.Grade)

  sens <- tp/(tp+fn)
```

```
  spec <- tn/(tn+fp)

  bacc[k] <- (sens+spec)/2
}

bacc

which(bacc==max(bacc))

pred.class <- as.numeric(pred >= which(bacc==max(bacc))/100)

confusionMatrix(factor(pred.class),
    factor(data.220$Course.Numeric.Grade), positive = '1')

roc.pred <- prediction(pred, data.220$Course.Numeric.Grade)

roc.perf <- performance(roc.pred, 'tpr', 'fpr')

plot(roc.perf, main='220 Model')
abline(0, 1, col='RED')

performance(roc.pred, 'auc')@y.values[[1]]

pred.test <- predict(mod, data.220, type = 'response')

pred.class.test <- as.numeric(pred.test >= which(bacc==max(bacc))/100)

accuracy.3 <- confusionMatrix(factor(pred.class.test),
    factor(data.220$Course.Numeric.Grade))$overall[1]
```