Investigating Bullying Factors in Adolescents

Kayla Ko, Tracy Huang, Cassie Jin

5/8/23

Contents

Abstract

This technical appendix investigates a dataset from the Global School-Based Student Health Survey (GSHS). The data consist of responses from a school-based, self-administered questionnaire assessing adolescents' demographics, family, and social factors related to bullying. We seek to answer the following questions: (1) which factors are associated with an adolescent being cyberbullied/bullied (2) which factors are associated with physical attack severity levels in bullying (3) can a classification model predict whether an adolescent was bullied based on various factors with a reasonable accuracy rate. Through logistic regression and ordinal logistic regression, we find that many factors, such as an adolescent's gender and loneliness level, are associated with in-person bullying, cyberbullying, and more frequent physical attacks. Through a classification tree model, we find that we can predict whether an adolescent is bullied or not based on information about loneliness and physical attacks, with an accuracy rate of 65.9%. We conclude with recommendations that schools and parents can implement to prevent or reduce bullying for adolescents.

Introduction

Bullying is a pervasive problem that negatively affects the wellbeing of students in all aspects of their lives. The 2019 National Center for Educational Statistics estimates that 1 out of 5 kids is bullied. Additionally, from the CDC's Youth Risk Behavior Surveillance System survey, 5.4 million kids stay at home at any given day in fear of being bullied. Thus, it is critical to better understand which factors are associated with bullying to create more targeted support systems and bullying prevention programs. This project aims to:

(1) Identify which factors are associated with an adolescent being cyberbullied/bullied, which will allow us to identify bullying risk factors that can be prevented or alleviated in adolescents.

(2) Understand which factors are associated with physical attack severity levels in bullying, to assist with identifying risk factors of being physically attacked for adolescents.

(3) Run a classification model to accurately predict whether an adolescent was bullied based on various factors, which can assist with predicting bullying in other student populations.

Dataset and Wrangling

The data set we used consists of data from the Global School-Based Student Health Survey (GSHS), a school-based survey which uses a self-administered questionnaire to obtain data on adolescent's demographics, family, and social factors related to bullying. The survey was conducted in Argentina in 2018 with 56,981 student participants.

After cleaning the data, there were 32,938 observations and twenty variables.

The variables in our dataset:

```
glimpse(data)
```

```
## Rows: 56,981
## Columns: 18
## $ record                                          <int> 1, 2, 3, 4, 5, 6, 7, ~
## $ bullied_on_school_property_in_past_12_months    <chr> "Yes", "No", "No", "N~
## $ bullied_not_on_school_property_in_past_12_months <chr> "Yes", "No", "No", "N~
## $ cyber_bullied_in_past_12_months                 <chr> " ", "No", "No", "No"~
## $ custom_age                                      <chr> "13 years old", "13 y~
## $ sex                                             <chr> "Female", "Female", "~
## $ physically_attacked                             <chr> "0 times", "0 times",~
## $ physical_fighting                               <chr> "0 times", "0 times",~
## $ felt_lonely                                     <chr> "Always", "Never", "N~
## $ close_friends                                   <chr> "2", "3 or more", "3 ~
## $ miss_school_no_permission                       <chr> "10 or more days", "0~
## $ other_students_kind_and_helpful                 <chr> "Never", "Sometimes",~
## $ parents_understand_problems                     <chr> "Always", "Always", "~
## $ most_of_the_time_or_always_felt_lonely          <chr> "Yes", "No", "No", "N~
## $ missed_classes_or_school_without_permission     <chr> "Yes", "No", "No", "N~
## $ were_underweight                                <chr> " ", " ", "No", "No",~
## $ were_overweight                                 <chr> " ", " ", "No", "No",~
## $ were_obese                                      <chr> " ", " ", "No", "No",~
```

Question 1: Which factors are associated with an adolescent being bullied?

Since bullying varies in degree and type, yet is still so pervasive, we will carry out an initial analysis of the data set that gives a glimpse into the patterns between different predictive factors and whether or not an adolescent is bullied. Every case of attack is different, but through finding relationships among the variables, we will gain a clearer understanding of the conditions that are often adjacent to being bullied. Although we may not draw any causal judgments on the presence of bullying, this analysis hopefully raises awareness on the concurrent signs of an adolescent suffering from external mental and physical attacks.

Data Processing and Methods

In order to determine which factors are associated with an adolescent being bullied, we will perform logistic regression analysis, which is used for predicting a binary variable with numerical or categorical data.

The data set records our dependent variable, occurrence of bullying, in three ways: whether an adolescent was bullied on school property (`Bullied_on_school_property_in_past_12_months`), whether an adolescent was bullied not on school property (`Bullied_not_on_school_property_in_past_12_months`), and whether an adolescent was cyberbullied (`Cyber_bullied_in_past_12_months`). All three variables were originally recorded as "Yes" or "No," which we recoded to 1 or 0, respectively. We then combined the first two variables into one denoting whether or not an adolescent was bullied in person, and we left the third variable as a count of cyberbullying.

- `isBulliedInPerson` - 1 if the adolescent was bullied on or off school property, 0 if the adolescent was neither bullied on school property nor off school property.
- `isCyberbullied` - 1 if the adolescent was cyberbullied, 0 if the adolescent was not cyberbullied.

All of the categorical variables in the dataset were recorded as strings containing ordinal values and words. For example, the categories of the number of days that an adolescent missed school without permission variable were "0 days," "1 to 2 days," "3 to 5 days," "6 to 9 days," and "10 or more." In order to make these data points usable as explanatory variables for logistic regression, we recoded them to clean out the words and assigned appropriate integer values to the groups. In the case of the days that an adolescent missed school without permission variable, "0 days" was sent to 0, "1 to 2 days" to 1.5, "3 to 5 days" to 4, "6 to 9 days" to 7.5, and "10 or more days" to 10. This allowed us to apply a simple but comprehensive understanding of the data when we later construct and interpret the final logistic regression model. These are the following predictors that we considered including in the final model:

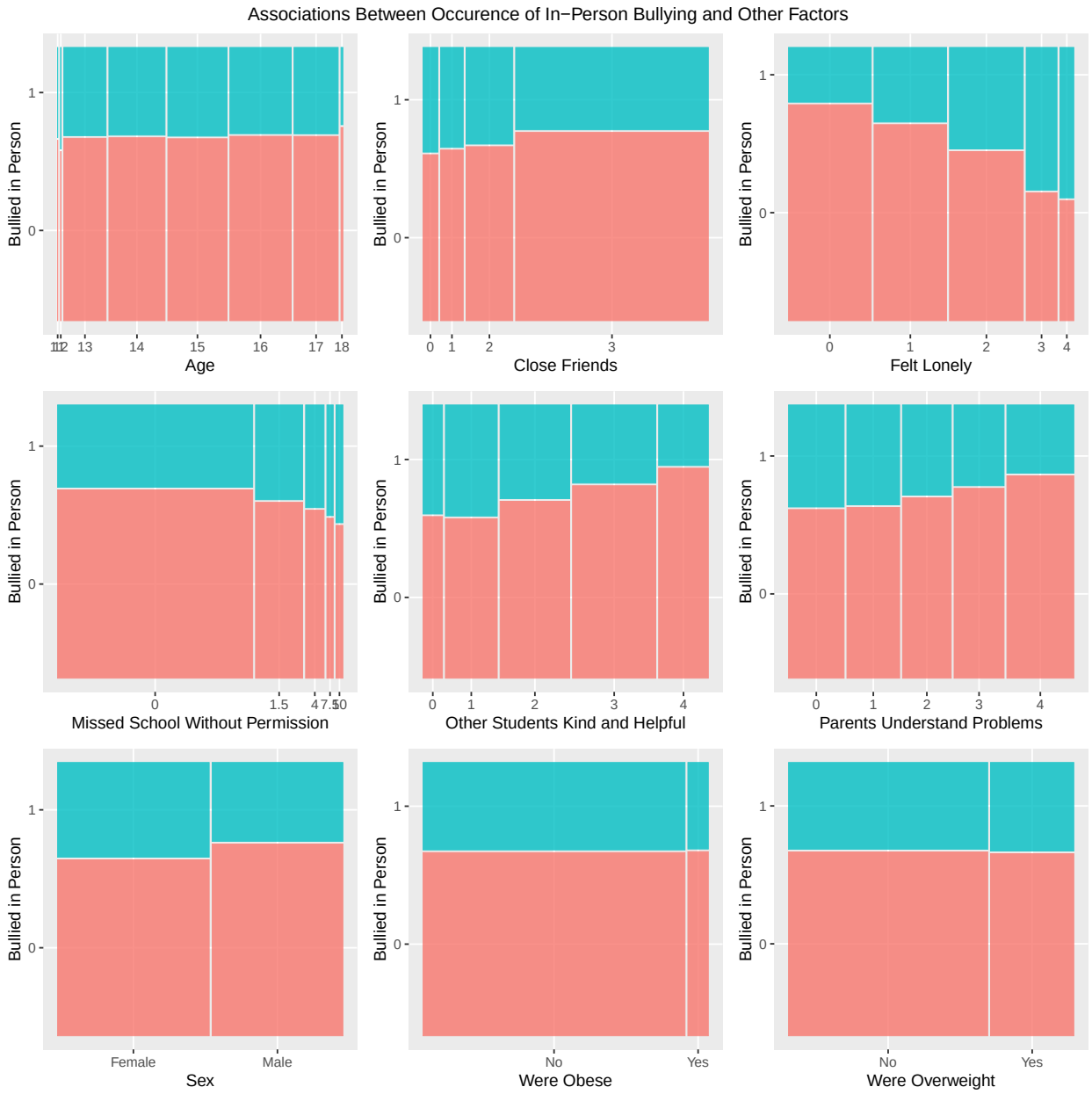
- `age` - Indicates an adolescent's age. Has integer values from 11 to 18
- `close_friends` - Indicates the number of close friends an adolescent has. Has integer values 0, 1, 2, and 3 if an adolescent has 3 or more friends
- `felt_lonely` - Indicates an adolescent's loneliness level. Has values 0, 1, 2, 3, and 4 for if an adolescent feels lonely never, rarely, sometimes, most of the time, or always, respectively
- `miss_school_no_permission` - Indicates how often an adolescent misses school without permission over the past 12 months. Has values 0, 1.5, 4, 7.5, 10 for if an adolescent misses 0 days, 1 to 2 days, 3 to 5 days, 6 to 9 days, or 10 or more days, respectively
- `other_students_kind_and_helpful` - Indicates how much an adolescent perceives other students at their school as kind and helpful. Has values 0, 1, 2, 3, and 4 for if an adolescent perceives other students at their school as kind and helpful never, rarely, sometimes, most of the time, or always, respectively

- `parents_understand_problems` - Indicates how much an adolescent believes their parents understand their problems. Has values 0, 1, 2, 3, and 4 for if an adolescent believes their parents understand their problems never, rarely, sometimes, most of the time, or always, respectively
- `sex` - Indicates an adolescent's gender. Has values "Female" and "Male"
- `were_obese` - Indicates an adolescent's obesity status. Has value "Yes" if the adolescent is obese, and "No" if the adolescent is not obese
- `were_overweight` - Indicates an adolescent's overweight status. Has value "Yes" if the adolescent is overweight, and "No" if the adolescent is not overweight

Exploratory Data Analysis

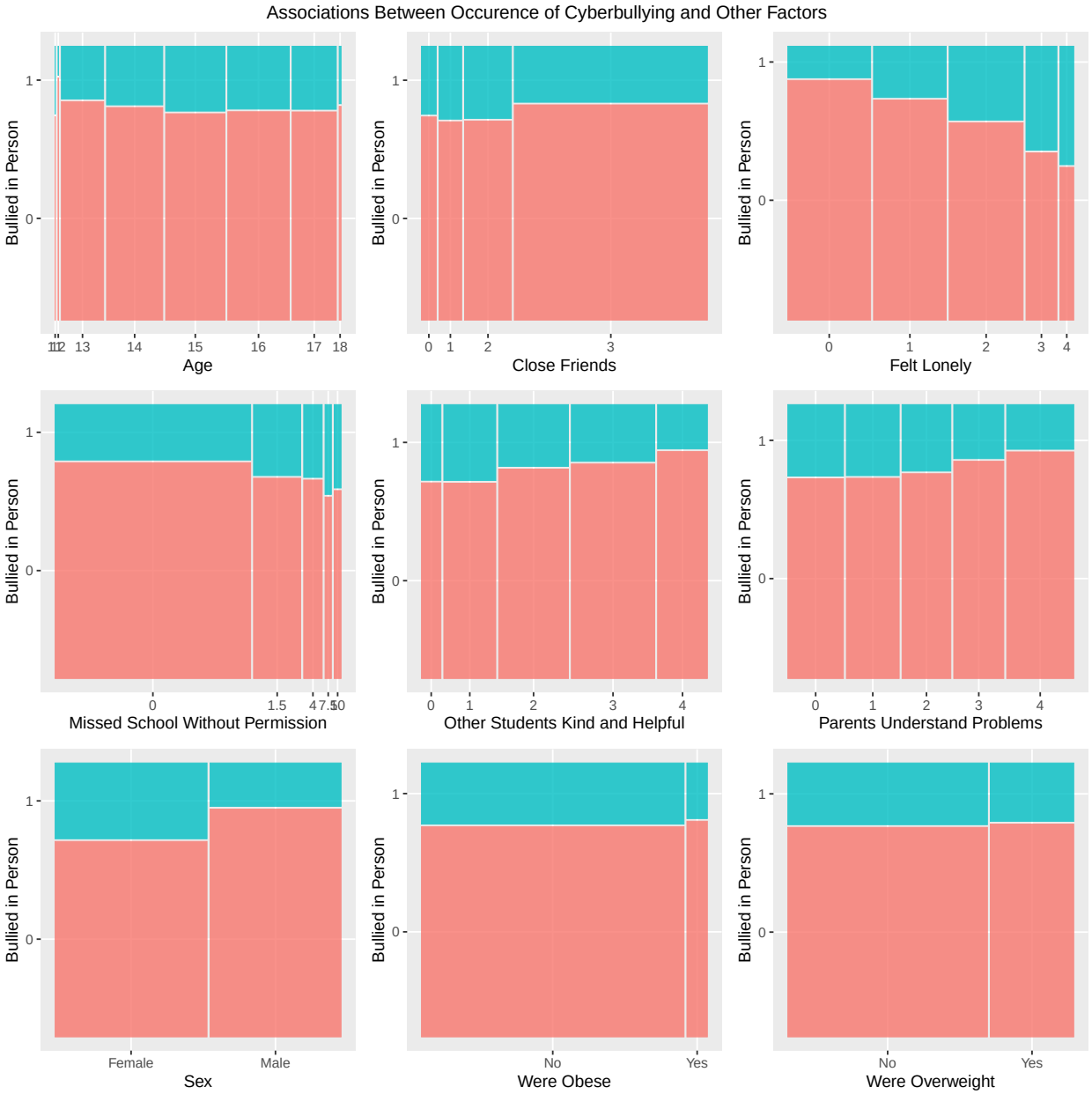
```
# in-person bullying
gg <- theme(legend.position = "none")

p1 <- ggplot(data = completedata_cleaned) +
  geom_mosaic(aes(x = product(isBulliedInPerson, age), fill=isBulliedInPerson)) +
  gg + xlab("Age") + ylab("Bullied in Person")

p2 <- ggplot(data = completedata_cleaned) +
  geom_mosaic(aes(x = product(isBulliedInPerson, close_friends), fill=isBulliedInPerson)) +
  gg + xlab("Close Friends") + ylab("Bullied in Person")

p3 <- ggplot(data = completedata_cleaned) +
  geom_mosaic(aes(x = product(isBulliedInPerson, felt_lonely), fill=isBulliedInPerson)) +
  gg + xlab("Felt Lonely") + ylab("Bullied in Person")

p4 <- ggplot(data = completedata_cleaned) +
  geom_mosaic(aes(x = product(isBulliedInPerson, miss_school_no_permission),
                  fill=isBulliedInPerson)) +
  gg + xlab("Missed School Without Permission") + ylab("Bullied in Person")

p5 <- ggplot(data = completedata_cleaned) +
  geom_mosaic(aes(x = product(isBulliedInPerson,
                              other_students_kind_and_helpful),
                  fill=isBulliedInPerson)) +
  gg + xlab("Other Students Kind and Helpful") + ylab("Bullied in Person")

p6 <- ggplot(data = completedata_cleaned) +
  geom_mosaic(aes(x = product(isBulliedInPerson, parents_understand_problems),
                  fill=isBulliedInPerson)) +
  gg + xlab("Parents Understand Problems") + ylab("Bullied in Person")

p7 <- ggplot(data = completedata_cleaned) +
  geom_mosaic(aes(x = product(isBulliedInPerson, sex), fill=isBulliedInPerson)) +
  gg + xlab("Sex") + ylab("Bullied in Person")

p8 <- ggplot(data = completedata_cleaned) +
  geom_mosaic(aes(x = product(isBulliedInPerson, were_obese), fill=isBulliedInPerson)) +
  gg + xlab("Were Obese") + ylab("Bullied in Person")
```

```
p9 <- ggplot(data = completedata_cleaned) +
  geom_mosaic(aes(x = product(isBulliedInPerson, were_overweight), fill=isBulliedInPerson)) +
  gg + xlab("Were Overweight") + ylab("Bullied in Person")

grid.arrange(grobs=list(p1,p2,p3,p4,p5,p6,p7,p8,p9), ncol = 3, nrow=3, common.legend=T,
             legend.position="bottom",
             top="Associations Between Occurence of In-Person Bullying and Other Factors")
```




```
# cyberbullying
gg <- theme(legend.position = "none")

p11 <- ggplot(data = completedata_cleaned) +
  geom_mosaic(aes(x = product(isCyberbullied, age), fill=isCyberbullied)) +
```

```
  gg + xlab("Age") + ylab("Bullied in Person")

p12 <- ggplot(data = completedata_cleaned) +
  geom_mosaic(aes(x = product(isCyberbullied, close_friends), fill=isCyberbullied)) +
  gg + xlab("Close Friends") + ylab("Bullied in Person")

p13 <- ggplot(data = completedata_cleaned) +
  geom_mosaic(aes(x = product(isCyberbullied, felt_lonely), fill=isCyberbullied)) +
  gg + xlab("Felt Lonely") + ylab("Bullied in Person")

p14 <- ggplot(data = completedata_cleaned) +
  geom_mosaic(aes(x = product(isCyberbullied, miss_school_no_permission), fill=isCyberbullied)) +
  gg + xlab("Missed School Without Permission") + ylab("Bullied in Person")

p15 <- ggplot(data = completedata_cleaned) +
  geom_mosaic(aes(x = product(isCyberbullied, other_students_kind_and_helpful), fill=isCyberbullied)) +
  gg + xlab("Other Students Kind and Helpful") + ylab("Bullied in Person")

p16 <- ggplot(data = completedata_cleaned) +
  geom_mosaic(aes(x = product(isCyberbullied, parents_understand_problems), fill=isCyberbullied)) +
  gg + xlab("Parents Understand Problems") + ylab("Bullied in Person")

p17 <- ggplot(data = completedata_cleaned) +
  geom_mosaic(aes(x = product(isCyberbullied, sex), fill=isCyberbullied)) +
  gg + xlab("Sex") + ylab("Bullied in Person")

p18 <- ggplot(data = completedata_cleaned) +
  geom_mosaic(aes(x = product(isCyberbullied, were_obese), fill=isCyberbullied)) +
  gg + xlab("Were Obese") + ylab("Bullied in Person")

p19 <- ggplot(data = completedata_cleaned) +
  geom_mosaic(aes(x = product(isCyberbullied, were_overweight), fill=isCyberbullied)) +
  gg + xlab("Were Overweight") + ylab("Bullied in Person")

grid.arrange(grobs=list(p11,p12,p13,p14,p15,p16,p17,p18,p19), ncol = 3, nrow=3, common.legend=T,
             legend.position="bottom",
             top="Associations Between Occurence of Cyberbullying and Other Factors")
```



In the mosaic plots displayed below, we see that there are differences in in-person and cyberbullying occurrence frequencies for the nine variables we considered in the data set.

- Older adolescents had just slightly lower proportions of experiencing in-person bullying and cyberbullying, compared to younger adolescents.
- Adolescents with a greater number of close friends had lower proportions of experiencing in-person bullying and cyberbullying, compared to adolescents with a smaller number of close friends.
- Adolescents who felt lonely more had drastically higher proportions of experiencing in-person bullying and cyberbullying, compared to adolescents who felt lonely less.
- Adolescents who missed school without permission more had increasingly higher proportions of experiencing in-person bullying and cyberbullying, compared to adolescents who did not miss school without permission as much.

- Adolescents who indicated that other students were more kind and helpful had lower proportions of experiencing in-person bullying and cyberbullying, compared to adolescents who indicated that other students were not as kind and helpful.
- Adolescents whose parents did not understand their problems had lower proportions of experiencing in-person bullying and cyberbullying, compared to adolescents whose parents understood their problems more.
- Male adolescents had lower proportions of experiencing in-person bullying and cyberbullying, compared to female adolescents.
- Obese adolescents had roughly the same proportions of experiencing in-person bullying and cyberbullying, compared to non-obese adolescents.
- Overweight adolescents had roughly the same proportions of experiencing in-person bullying and cyberbullying, compared to non-obese adolescents.

Model Fitting

Since the conditions for logistic regression are satisfied, we proceed to fit a logistic regression model:

```
# fit model
logmod_inperson <- glm(isBulliedInPerson ~ age + close_friends + felt_lonely +
                         miss_school_no_permission +
                         other_students_kind_and_helpful +
                         parents_understand_problems + sex + were_obese +
                         were_overweight,
                       data = completedata_cleaned, family = binomial)

# obtain model summary
summary(logmod_inperson)
```

```
##
## Call:
## glm(formula = isBulliedInPerson ~ age + close_friends + felt_lonely +
##     miss_school_no_permission + other_students_kind_and_helpful +
##     parents_understand_problems + sex + were_obese + were_overweight,
##     family = binomial, data = completedata_cleaned)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.7315  -0.8757  -0.6976   1.2211   1.9990
##
## Coefficients:
##                                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)                     -0.103723   0.148327  -0.699   0.4844
## age                             -0.059564   0.009238  -6.448 1.13e-10 ***
## close_friends                    0.007334   0.014484   0.506   0.6126
## felt_lonely                      0.414608   0.011633  35.640  < 2e-16 ***
## miss_school_no_permission        0.050746   0.005692   8.916  < 2e-16 ***
## other_students_kind_and_helpful -0.166010   0.010836 -15.320  < 2e-16 ***
## parents_understand_problems     -0.020585   0.008936  -2.304   0.0212 *
## sexMale                          0.003100   0.025754   0.120   0.9042
## were_obeseYes                   -0.065643   0.052455  -1.251   0.2108
```

```
## were_overweightYes               0.055602   0.029788   1.867   0.0620 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 41437  on 32937  degrees of freedom
## Residual deviance: 39112  on 32928  degrees of freedom
## AIC: 39132
##
## Number of Fisher Scoring iterations: 4
```

We observe that most of the coefficients are statistically significant for predicting the occurrence of in-person bullying, as their p-values are less than $\alpha = 0.05$. Namely, an adolescent's age, level of loneliness, frequency of missing school without permission, perception of other students as kind and helpful, and level of understanding from parents are associated with being bullied in person. Four of the predictors yield insignificant coefficients (p-values greater than $\alpha = 0.05$): the number of close friends an adolescent has, whether or not the adolescent is male, whether or not the adolescent is obese, and whether or not the adolescent is overweight.

```
# fit model
logmod_extra1 <- glm(isBulliedInPerson ~ close_friends + sex + were_obese + were_overweight,
                     data = completedata_cleaned, family = binomial)

# obtain model summary
summary(logmod_extra1)
```

```
##
## Call:
## glm(formula = isBulliedInPerson ~ close_friends + sex + were_obese +
##     were_overweight, family = binomial, data = completedata_cleaned)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.0757  -0.8954  -0.8127   1.4285   1.6059
##
## Coefficients:
##                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)        -0.30551    0.03701  -8.255   <2e-16 ***
## close_friends      -0.13386    0.01345  -9.954   <2e-16 ***
## sexMale            -0.26018    0.02398 -10.849   <2e-16 ***
## were_obeseYes      -0.03239    0.05066  -0.639   0.5226
## were_overweightYes  0.06141    0.02873   2.138   0.0325 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 41437  on 32937  degrees of freedom
## Residual deviance: 41205  on 32933  degrees of freedom
## AIC: 41215
##
## Number of Fisher Scoring iterations: 4
```

Fitting a model for occurrence of in-person bullying with just the four variables that proved statistically insignificant above, we see that three of the four predictors are significant in a smaller model: the number of close friends an adolescent has, whether or not the adolescent is male, and whether or not the adolescent is overweight. This may be due to a few reasons: loss of degrees of freedom, multi-collinearity, and/or misspecified models, where regressing with fewer predictors increases the possibility that the small model suffers from omitted variable bias. We see in the next section of conditions that the model satisfies the multi-collinearity assumption. However, for the sake of parsimony, we proceed without these variables, with caution.

```
# fit model
logmod_inperson_final <- glm(isBulliedInPerson ~ age + felt_lonely +
                               miss_school_no_permission +
                               other_students_kind_and_helpful +
                               parents_understand_problems,
                             data = completedata_cleaned, family = binomial)

# obtain model summary
summary(logmod_inperson_final)
```

```
##
## Call:
## glm(formula = isBulliedInPerson ~ age + felt_lonely + miss_school_no_permission +
##     other_students_kind_and_helpful + parents_understand_problems,
##     family = binomial, data = completedata_cleaned)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.7087  -0.8758  -0.6957   1.2204   1.9894
##
## Coefficients:
##                                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)                     -0.052104   0.140858  -0.370   0.7115
## age                             -0.060898   0.009186  -6.629 3.37e-11 ***
## felt_lonely                      0.413012   0.011074  37.295  < 2e-16 ***
## miss_school_no_permission        0.050908   0.005685   8.955  < 2e-16 ***
## other_students_kind_and_helpful -0.165237   0.010735 -15.393  < 2e-16 ***
## parents_understand_problems     -0.020467   0.008921  -2.294   0.0218 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 41437  on 32937  degrees of freedom
## Residual deviance: 39116  on 32932  degrees of freedom
## AIC: 39128
##
## Number of Fisher Scoring iterations: 4
```

```
# fit model
logmod_cyber <- glm(isCyberbullied ~ age + close_friends + felt_lonely +
                      miss_school_no_permission +
                      other_students_kind_and_helpful +
                      parents_understand_problems + sex + were_obese +
```

```
                      were_overweight,
                    data = completedata_cleaned, family = binomial)

# obtain model summary
summary(logmod_cyber)
```

```
##
## Call:
## glm(formula = isCyberbullied ~ age + close_friends + felt_lonely +
##     miss_school_no_permission + other_students_kind_and_helpful +
##     parents_understand_problems + sex + were_obese + were_overweight,
##     family = binomial, data = completedata_cleaned)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.5099  -0.7394  -0.5665  -0.4250   2.2563
##
## Coefficients:
##                                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)                     -1.587051   0.166895  -9.509  < 2e-16 ***
## age                              0.008524   0.010349   0.824  0.41017
## close_friends                    0.013757   0.015999   0.860  0.38986
## felt_lonely                      0.390047   0.012753  30.584  < 2e-16 ***
## miss_school_no_permission        0.059093   0.006102   9.684  < 2e-16 ***
## other_students_kind_and_helpful -0.095078   0.012132  -7.837 4.61e-15 ***
## parents_understand_problems     -0.028216   0.010055  -2.806  0.00501 **
## sexMale                         -0.476128   0.029344 -16.225  < 2e-16 ***
## were_obeseYes                   -0.075514   0.060306  -1.252  0.21050
## were_overweightYes               0.012679   0.033645   0.377  0.70629
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 35046  on 32937  degrees of freedom
## Residual deviance: 32846  on 32928  degrees of freedom
## AIC: 32866
##
## Number of Fisher Scoring iterations: 4
```

Again, we observe that most of the coefficients are significant. There is evidence that an adolescent's level of loneliness, frequency of missing school without permission, perception of other students as kind and helpful, level of understanding from parents, and an adolescent being male, are associated with being cyberbullied. On the other hand, an adolescent's age, the number of close friends an adolescent has, whether or not the adolescent is obese, and whether or not the adolescent is overweight, are not significant predictors of occurrence of cyberbullying.

```
# fit model
logmod_extra2 <- glm(isCyberbullied ~ age + close_friends + were_obese +
                       were_overweight,
                     data = completedata_cleaned, family = binomial)
```

```r
# obtain model summary
summary(logmod_extra2)
```

```
##
## Call:
## glm(formula = isCyberbullied ~ age + close_friends + were_obese +
##     were_overweight, family = binomial, data = completedata_cleaned)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.8756  -0.7218  -0.6949  -0.6556   1.8889
##
## Coefficients:
##                     Estimate Std. Error z value Pr(>|z|)
## (Intercept)        -1.534310   0.156909  -9.778  < 2e-16 ***
## age                 0.042950   0.009861   4.355 1.33e-05 ***
## close_friends      -0.136015   0.014757  -9.217  < 2e-16 ***
## were_obeseYes      -0.090481   0.058156  -1.556    0.120
## were_overweightYes -0.039732   0.032446  -1.225    0.221
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 35046  on 32937  degrees of freedom
## Residual deviance: 34931  on 32933  degrees of freedom
## AIC: 34941
##
## Number of Fisher Scoring iterations: 4
```

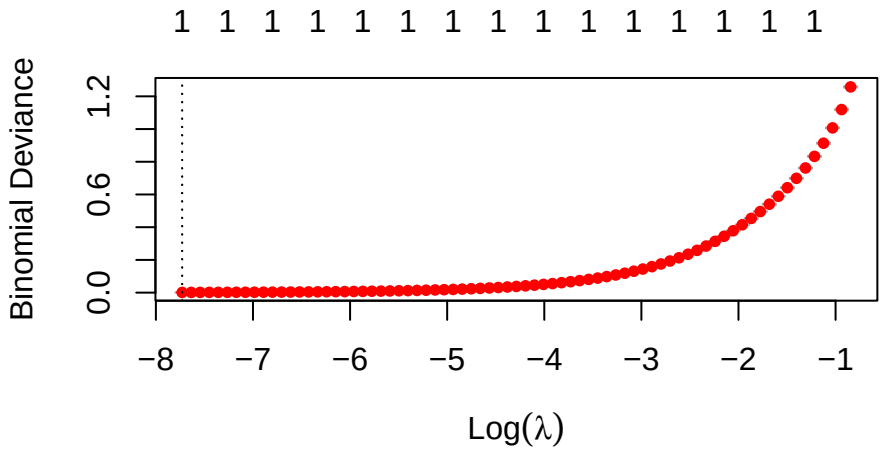
We perform one more logistic regression to determine whether or not the insignificant variables in the large model for predicting occurrence of cyberbullying are in fact significant on a smaller scale. According to the p-values, the more specific model yields significant coefficient values for the variables age and number of close friends, but the variables that capture whether or not an adolescent is obese and whether or not an adolescent is overweight are still insignificant. Thus, there is no evidence of significant associations between an adolescent's weight level and occurrence of cyberbullying and we remove these two predictors from the model from this point onward. Like the logistic regression model for in person bullying, we safely proceed with a simpler model that does not include any of these four predictors.

```
##
## Call:
## glm(formula = isCyberbullied ~ felt_lonely + miss_school_no_permission +
##     other_students_kind_and_helpful + parents_understand_problems +
##     sex, family = binomial, data = completedata_cleaned)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.4954  -0.7383  -0.5638  -0.4242   2.2265
##
## Coefficients:
##                                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)                    -1.428865   0.046321 -30.847  < 2e-16 ***
## felt_lonely                     0.388975   0.012550  30.994  < 2e-16 ***
```

```
## miss_school_no_permission        0.059520   0.006062   9.819  < 2e-16 ***
## other_students_kind_and_helpful -0.093181   0.012007  -7.761 8.45e-15 ***
## parents_understand_problems     -0.028182   0.010039  -2.807  0.00499 **
## sexMale                         -0.476766   0.029193 -16.331  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 35046  on 32937  degrees of freedom
## Residual deviance: 32849  on 32932  degrees of freedom
## AIC: 32861
##
## Number of Fisher Scoring iterations: 4
```

```
# store summaries in table
(clogmod1 <- coef(summary(logmod_inperson_final)))
(clogmod2 <- coef(summary(logmod_cyber_final)))
```

```
# logistic regression with lasso
full_data <- completedata_cleaned %>%
 dplyr::relocate(isBulliedInPerson)

x_data <- model.matrix(isBulliedInPerson ~ ., full_data)[,-1]
y_data <- full_data %>%
 dplyr::select(isBulliedInPerson) %>%
 unlist() %>%
 as.numeric()
lasso_model <- cv.glmnet(x_data, y_data, family = "binomial", alpha = 1)

plot(lasso_model)
```




```
coef(lasso_model, lasso_model$lambda.1se)
```

```
## 24 x 1 sparse Matrix of class "dgCMatrix"
##                                                                    s1
## (Intercept)                                                -7.726551
## miss_school_no_permission                                   .
## physical_fighting                                           .
## physically_attacked                                         .
## age                                                         .
## record                                                      .
## bullied_on_school_property_in_past_12_months                .
## bullied_not_on_school_property_in_past_12_months            .
## cyber_bullied_in_past_12_months                             .
## sexMale                                                     .
## felt_lonely                                                 .
## close_friends                                               .
## other_students_kind_and_helpful                             .
## parents_understand_problems                                 .
## most_of_the_time_or_always_felt_lonelyYes                   .
## missed_classes_or_school_without_permissionYes              .
## were_underweightYes                                         .
## were_overweightYes                                          .
## were_obeseYes                                               .
## `bullied_on_school_property_in_past_12_months <- ...`       .
## `bullied_not_on_school_property_in_past_12_months <- ...`   .
## `cyber_bullied_in_past_12_months <- ...`                    .
## total                                                      14.305696
## isCyberbullied1                                             .
```

In the case that none of our predictors exhibit significant associations with our response variables, occurrence of in-person bullying and cyberbullying, we use glmnet, a package that fits a generalized logistic model to our data via penalized maximum likelihood. The package includes methods for prediction and plotting, and functions for cross-validation. By producing a relaxed lasso regression model between all of the independent variables and the dependent ones, we obtain a more lenient regression standard. The coefficients from this analysis are λ values, which are hyperparameters that capture many layers of information between and among the variables. The best lambda minimizes error in prediction, and we see from the coefficient values that, in fact, none of our predictor variables are significant in the relaxed lasso regression. Thus, we decided to proceed with the more stringent method of the above standard logistic regression.

Checking Model Assumptions

1. The dependent variable is binary.

Our dependent variables, occurrence of bullying in person and occurrence of cyberbullying, are both binary, so this assumption is satisfied.

2. The observations are independent of each other.

The observations in the data set were independently collected and did not come from repeated measurements or matched data. Thus, we have independent observations.

3. There is no multi-collinearity among the independent variables.

```
vif(logmod_inperson_final)
```

```
##                             age                     felt_lonely
##                        1.027505                        1.108710
##       miss_school_no_permission other_students_kind_and_helpful
##                        1.018812                        1.062475
##     parents_understand_problems
##                        1.132493
```

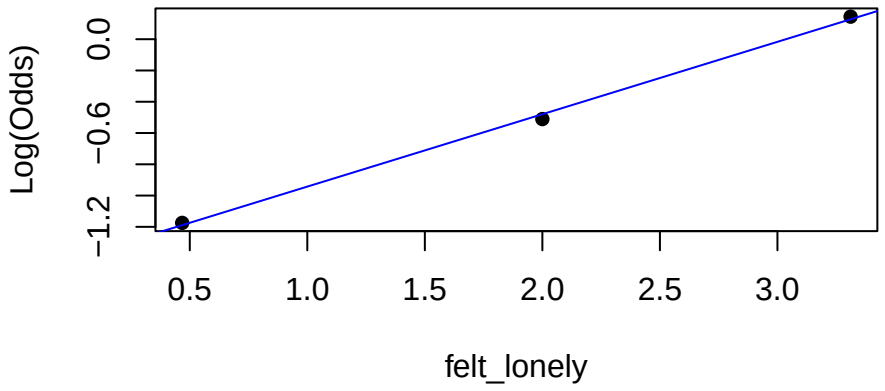
```
vif(logmod_cyber_final)
```

```
##                     felt_lonely       miss_school_no_permission
##                        1.174874                        1.009514
## other_students_kind_and_helpful     parents_understand_problems
##                        1.070083                        1.138535
##                             sex
##                        1.058396
```

The VIF score for all of the independent variables is around 1. Since they are all less than 5, the independent variables are not too highly correlated with each other and we do not have issues of multi-collinearity.

4. The independent variables and log odds are linear.

```
emplogitplot1(isBulliedInPerson ~ felt_lonely, data = completedata_cleaned)
```



We see from the plot of the logit function vs. one of our predictor variables, level of loneliness, that there is a linear pattern between the two factors. This trend persists across all of the significant predictors, thus we have that the logit transformation of probabilities is a linear function of the predictors.

5. The sample size is large.

Since there are more than 10 observations of the least frequent outcome for each independent variable in our model, the sample size is sufficiently large.

Model Interpretation

```
# obtain odds ratios for in-person bullying
exp(coef(logmod_inperson_final))
```

- Age: For every level increase in age, the odds of an adolescent being more likely to be bullied in person becomes smaller by 0.94 times, holding constant all other variables.
- Felt Lonely: For every level increase in how lonely an adolescent feels, the odds of being more likely to be bullied in person becomes larger by 1.51 times, holding constant all other variables.
- Missed School Without Permission: For every level increase in frequency of an adolescent missing school without permission, the odds of being more likely to be bullied in person becomes larger by 1.05 times, holding constant all other variables.
- Other Students' Kindness and Helpfulness: For every level increase in how much an adolescent finds other students to be kind and helpful, the odds of being more likely to be bullied in person becomes smaller by 0.85 times, holding constant all other variables.
- Parents Understanding of Problems: For every level increase in how much an adolescent feels their parents understand their problems, the odds of being more likely to be bullied in person becomes smaller by 0.98 times, holding constant all other variables.

```
# obtain odds ratios for cyberbullying
exp(coef(logmod_cyber_final))
```

- Felt Lonely: For every level increase in how lonely an adolescent feels, the odds of being more likely to be cyberbullied becomes larger by 1.48 times, holding constant all other variables.
- Missed School Without Permission: For every level increase in frequency of an adolescent missing school without permission, the odds of being more likely to be cyberbullied becomes larger by 1.06 times, holding constant all other variables.
- Other Students' Kindness and Helpfulness: For every level increase in how much an adolescent finds other students to be kind and helpful, the odds of being more likely to be cyberbullied becomes smaller by 0.91 times, holding constant all other variables.
- Parents Understanding of Problems: For every level increase in how much an adolescent feels their parents understand their problems, the odds of being more likely to be cyberbullied becomes smaller by 0.97 times, holding constant all other variables.
- Sex: For male adolescents, the odds of being more likely to be cyberbullied is 0.62 times that of female adolescents, holding constant all other variables.

Model Diagnostics

```
# Pseudo R2 for in-person bullying
DescTools::PseudoR2(
  logmod_inperson_final,
  which = c("McFadden"))
```

```
##  McFadden
## 0.0560164
```

```r
# Pseudo R2 for cyberbullying
DescTools::PseudoR2(
  logmod_cyber_final,
  which = c("McFadden"))
```

```
##   McFadden
## 0.06269198
```

Since logistical regression modeling is not linear, we may not use the coefficient of determination, R-squared as a measure for the model's goodness of fit. Instead, we implement a pseudo R-squared method for our logistic regression model. A McFadden's pseudo R-squared value between 0.2 and 0.4 indicates excellent model fit. However, based on the McFadden's pseudo R-squared value of approximately 0.06 for both our in-person bullying and cyberbullying models, we conclude that our model has weak predictive capability.

Key Takeaways

1. The general regression model to predict physical attack severity levels was valid. After removing a few predictors, all of the coefficients were statistically significant.
2. There are significant associations between an adolescent experiencing in-person bullying or cyberbullying, and being younger, feeling more lonely, missing school without permission more frequently, finding other students to be kind and helpful less, and feeling their parents understand them less.

Question 2: Which factors specifically affect the severity of physical attacks in bullying?

Bullying often has many different dimensions, such as isolation, rumor-spreading, and physical attacks. The data set unfortunately did not contain many variables measuring the specific aspects of bullying. Instead, the data set includes a variable measuring the number of physical attacks, so we wanted to investigate which factors affect the severity of physical attacks in bullying.

Beyond the injuries from physical attacks, the ongoing stress and trauma can also lead to other physical problems like sleep disorders, headaches, and chronic pain. Additionally, physical attacks also negatively affect adolescents' mental health, biological stress system, and academics. Thus, it is critical to better understand which factors are related to not just bullying but physical attacks specifically.

Data Processing and Methods

In order to predict the severity of physical attacks in bullying, we will use an ordinal logistic regression model, which is used for predicting an ordinal variable with more than two levels. Ordinal variables contain categories in a specific order.

The dependent variable is the severity of physical attacks. The physical attack variable (`physically_attacked`) was originally recorded in different categories based on frequency (0 times, 1 times, 2 to 3 times, etc.). However, we recoded the variable to denote different severity levels (low, medium, or high) based on the frequency.

- `physical_attack_severity` - Recoded to have values "Low" if the adolescent was physically attacked 0 or 1 times, "Medium" if the adolescent was physically attacked 2 to 7 times, or "High" if the adolescent was physically attacked 8 or more times.

Additionally, categorical variables with more than two levels were recoded to have only two levels. This was so the results would be easier to interpret in the final ordinal logistic regression model. These are the following predictors that we considered including in the final model because their levels' proportions differed based on the physical attack severity levels in mosaic plots (shown below in EDA section):

- `were_obese` - Has values "Yes" or "No"
- `most_of_the_time_or_always_felt_lonely` - Has values "Yes" or "No"
- `sex` - Has values "Male" or "Female"
- `other_students_kind_and_helpful` - Recoded to have values "Yes" if adolescents indicate that other students are kind and helpful always or most of the time, or "No" if adolescents indicate other students are kind and helpful sometimes, rarely, or never
- `parents_understand_problems` - Recoded to have values "Yes" if adolescents indicate that parents understand their problems always or most of the time, or "No" if adolescents indicate that parents understand their problems sometimes, rarely, or never
- `close_friends` - Recoded to have values "Low" if an adolescent has 0, 1, or 2 close friends, and "High" if an adolescent has 3 or more friends

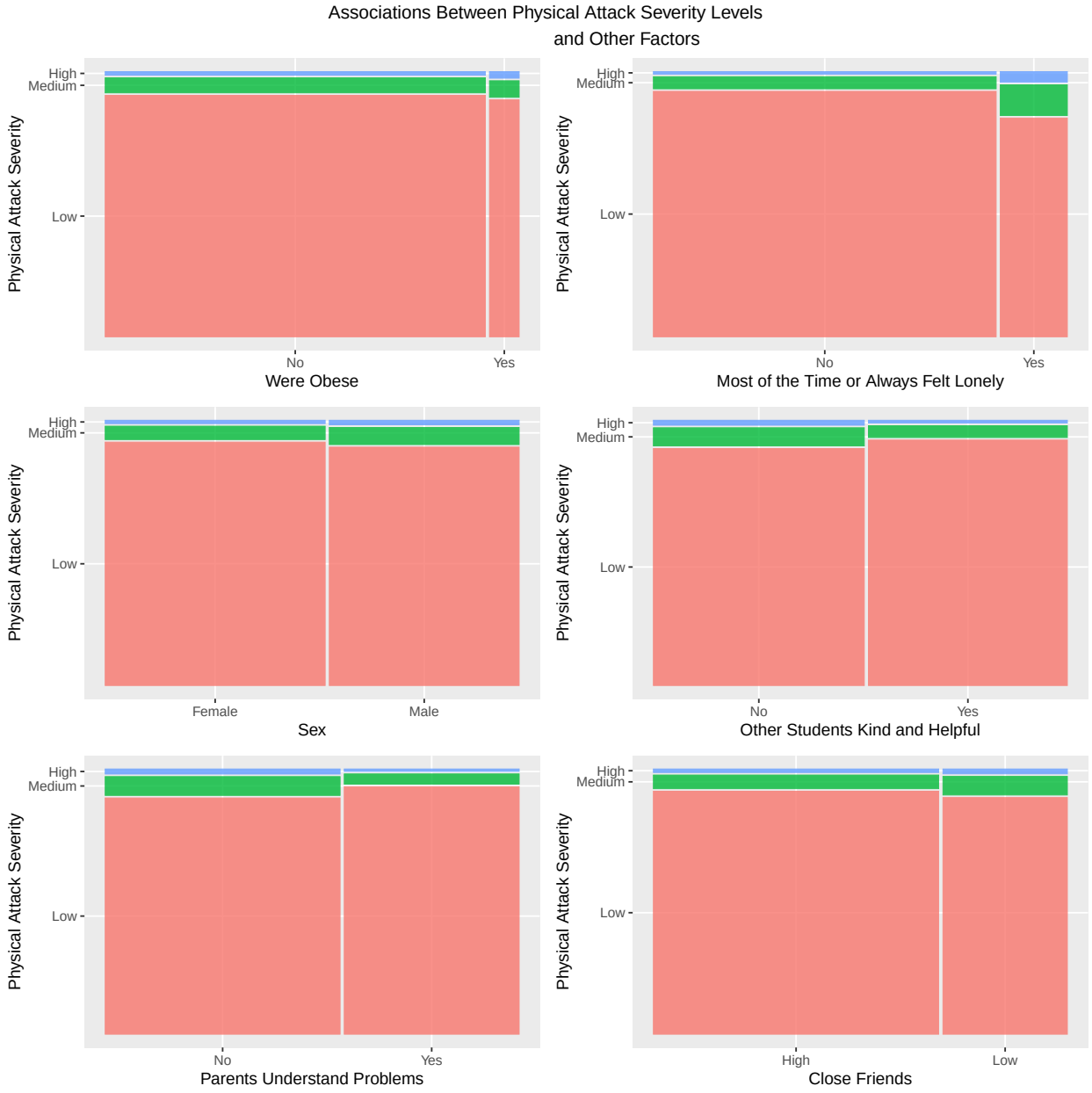
Exploratory Data Analysis

In the mosaic plots displayed below, we can see that there are differences in physical attack severity levels for the six variables described in the previous section.

- Obese adolescents had higher proportions of experiencing medium or high levels of physical attacks, compared to non-obese adolescents.
- Adolescents who felt lonely most of the time or always had higher proportions of experiencing medium or high levels of physical attacks, compared to adolescents who did not feel lonely most of the time or always.
- Male adolescents had higher proportions of experiencing medium or high levels of physical attacks, compared to female adolescents.
- Adolescents who indicated that other students were not kind and helpful had higher proportions of experiencing medium or high levels of physical attacks, compared to adolescents who indicated that other students were kind and helpful.
- Adolescents whose parents did not understand their problems had higher proportions of experiencing medium or high levels of physical attacks, compared to adolescents whose parents understood their problems.
- Adolescents with a smaller number of close friends had higher proportions of experiencing medium or high levels of physical attacks, compared to adolescents with a larger number of close friends.

```
gg <- theme(legend.position = "none")

p1 <- ggplot(data = physical_attack_data) +
  geom_mosaic(aes(x = product(physical_attack_severity, were_obese),
                  fill=physical_attack_severity)) +
  gg + xlab("Were Obese") + ylab("Physical Attack Severity")

p2 <- ggplot(data = physical_attack_data) +
  geom_mosaic(aes(x = product(physical_attack_severity, most_of_the_time_or_always_felt_lonely),
                  fill=physical_attack_severity)) +
  gg + xlab("Most of the Time or Always Felt Lonely") + ylab("Physical Attack Severity")

p3 <- ggplot(data = physical_attack_data) +
  geom_mosaic(aes(x = product(physical_attack_severity, sex),
                  fill=physical_attack_severity)) +
  gg + xlab("Sex") + ylab("Physical Attack Severity")

p4 <- ggplot(data = physical_attack_data) +
  geom_mosaic(aes(x = product(physical_attack_severity, other_students_kind_and_helpful),
                  fill=physical_attack_severity)) +
  gg + xlab("Other Students Kind and Helpful") + ylab("Physical Attack Severity")

p5 <- ggplot(data = physical_attack_data) +
  geom_mosaic(aes(x = product(physical_attack_severity, parents_understand_problems),
                  fill=physical_attack_severity)) +
  gg + xlab("Parents Understand Problems") + ylab("Physical Attack Severity")

p6 <- ggplot(data = physical_attack_data) +
  geom_mosaic(aes(x = product(physical_attack_severity, close_friends),
```

```
                  fill=physical_attack_severity)) +
  gg + xlab("Close Friends") + ylab("Physical Attack Severity")

grid.arrange(grobs=list(p1,p2,p3,p4,p5,p6), ncol = 2, nrow=3, common.legend=T,
legend.position="bottom", top="Associations Between Physical Attack Severity Levels
                               and Other Factors")
```



Model Fitting

Before fitting an ordinal logistic regression model, one assumption for the model that needs to be met is the proportional odds assumption. The assumption states that the coefficients that describe the relationship between "Low" physical attacks and "Medium" and "High" physical attacks are the same as those that

describe the relationship between "Medium" and "High" physical attacks. Thus, out of the predictors we were considering based on the mosaic plots, we first examined which variables violate the proportional odds assumption by running a nominal test on the ordinal regression model.

```
model <- clm(as.factor(physical_attack_severity) ~ were_obese +
               most_of_the_time_or_always_felt_lonely + sex +
               other_students_kind_and_helpful + parents_understand_problems +
               close_friends, data = physical_attack_data)
nominal_test(model)
```

```
## Tests of nominal effects
##
## formula: as.factor(physical_attack_severity) ~ were_obese + most_of_the_time_or_always_felt_lonely +
##                                       Df  logLik   AIC     LRT  Pr(>Chi)
## <none>                                   -9427.3 18871
## were_obese                             1 -9423.2 18865  8.0911 0.0044482 **
## most_of_the_time_or_always_felt_lonely  1 -9420.7 18860 13.1248 0.0002914 ***
## sex                                    1 -9427.2 18872  0.1950 0.6587648
## other_students_kind_and_helpful        1 -9427.0 18872  0.5612 0.4537793
## parents_understand_problems            1 -9427.1 18872  0.4516 0.5015731
## close_friends                          1 -9427.3 18873  0.0489 0.8250277
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

A low p-value in the nominal test indicates that the coefficient does not satisfy the proportional odds assumption. From the nominal test results, we see that the variables `were_obese` and `most_of_the_time_or_always_felt_lonely` have p-values less than 0.05, which means that they violate the proportional odds assumption. Thus, they will be excluded in the final ordinal logistic regression model.

We then fit the final ordinal logistic regression model:

```
# fit model
m <- polr(as.factor(physical_attack_severity) ~ sex + other_students_kind_and_helpful +
            parents_understand_problems + close_friends, data = physical_attack_data,
            Hess = TRUE)
```

```
# store summary in table
(ctable <- coef(summary(m)))
```

```
# obtain p-values
p <- pnorm(abs(ctable[, "t value"]), lower.tail = FALSE) * 2

# combined table with summary and p-values
(ctable <- cbind(ctable, "p value" = p))
```

```
##                                       Value Std. Error    t value
## sexMale                           0.3472986 0.04276595   8.120914
## other_students_kind_and_helpfulYes -0.3593484 0.04476173  -8.028027
## parents_understand_problemsYes    -0.6044809 0.04766091 -12.682949
## close_friendsLow                   0.2537049 0.04478157   5.665385
## Low|Medium                         2.4199498 0.04184805  57.827065
## Medium|High                        4.0707541 0.05781877  70.405409
##                                       p value
```

```
## sexMale                              4.626877e-16
## other_students_kind_and_helpfulYes 9.905282e-16
## parents_understand_problemsYes     7.351322e-37
## close_friendsLow                   1.466944e-08
## Low|Medium                         0.000000e+00
## Medium|High                        0.000000e+00
```

The regression output "Coefficients" table includes the values of each coefficient, standard errors, and t-values, which are the ratios of the coefficients to their standard error, and p-values. We observe that all of the coefficients are statistically significant, since their p-values are less than 0.05.

The intercepts table contains estimates for the two intercepts. The intercepts show where the latent or underlying variable is cut to make the three groups (Low, Medium, High) we observe in the physical attacks data. However, these intercepts are generally not used in the interpretation of results.

Lastly, the residual deviance and AIC values are presented at the bottom of the table. Residual deviance is a measure of how well the dependent variable is predicted with the model with only an intercept. AIC is a measure used to estimate prediction error and compare different regression models.

Model Interpretation

The coefficients in the output table are difficult to interpret because they are scaled in terms of logs. To understand them, it is easier to convert the coefficients into odds ratios. Thus, we calculated the odds ratios and their confidence intervals by exponentiating the estimates and their confidence intervals.

```
# obtain confidence intervals
(ci <- confint(m))
```

```
## Waiting for profiling to be done...
```

```
## obtain odds ratios
exp(coef(m))
```

```
## combine OR and CI into one table
exp(cbind(OR = coef(m), ci))
```

```
##                                        OR     2.5 %    97.5 %
## sexMale                              1.4152392 1.3015187 1.5390876
## other_students_kind_and_helpfulYes 0.6981311 0.6393299 0.7619676
## parents_understand_problemsYes     0.5463580 0.4973707 0.5995570
## close_friendsLow                   1.2887914 1.1801784 1.4066673
```

The odds ratios can be interpreted as follows:

- Sex: For male adolescents, the odds of being more likely to be physically attacked is 1.42 times that of female adolescents, holding constant all other variables.
- Other Students' Kindness and Helpfulness: For students who identified other students as not being kind and helpful, the odds of being more likely to be physically attacked is 1.43 times that of students who identified other students as being kind and helpful, holding constant all other variables.
- Parents Understanding of Problems: For students whose parents do not understand their problems, the odds of being more likely to be physically attacked is 1.83 times that of students whose parents do understand their problems, holding constant all other variables.

- Close Friends: For students who do not have many close friends, the odds of being more likely to be physically attacked is 1.29 times that of students who have many close friends, holding constant all other variables.

Checking Model Assumptions

1. The dependent variable is ordered.

This assumption is met because the dependent variable, `Physical_attack_severity`, is ordered, as it goes from low to medium to high.

2. One or more of the independent variables are either continuous, categorical, or ordinal.

This assumption is met because the independent variables are all categorical or ordinal.

3. There is no multi-collinearity amongst the variables.

```
vif(m)
```

```
##                            sex other_students_kind_and_helpful
##                       1.010945                         1.049288
##     parents_understand_problems                    close_friends
##                       1.041269                         1.032600
```

The VIF score for all of the independent variables is around 1. The general rule of thumb for the VIF test is that if the VIF value is greater than 5, there is multi-collinearity. However, since none of the VIF scores are greater than 5, there are no multi-collinearity issues in the model and the multi-collinearity assumption is met.

4. The model meets the proportional odds assumption.

```
brant(m)
```

```
## ------------------------------------------------------------------
## Test for             X2  df  probability
## ------------------------------------------------------------------
## Omnibus                  1.31    4   0.86
## sexMale                  0.15    1   0.7
## other_students_kind_and_helpfulYes   0.58    1   0.45
## parents_understand_problemsYes   0.41    1   0.52
## close_friendsLow         0.04    1   0.84
## ------------------------------------------------------------------
##
## H0: Parallel Regression Assumption holds
```

The Brant Test is used for testing whether an ordinal logistic regression model violates the proportional odds assumption or not. A low p-value in the Brant test indicates that the coefficient does not satisfy the proportional odds assumption. However, since all of the coefficients for the predictors are much greater than 0.05, this means that they satisfy the proportional odds assumption. The output also includes an Omnibus variable that stands for the whole model; since its p-value is also much greater than 0.05, this further confirms that the model is valid and the proportional odds assumption is not violated.

Model Diagnostics

```
# pseudo R2
DescTools::PseudoR2(
  m, which = c("McFadden"))
```

```
##   McFadden
## 0.02011633
```

Like the previous section of logistic regression analysis, we utilized a pseudo R-squared method to obtain a measure for the model's goodness of fit, or how well the observed data corresponds to the model. Based on the low McFadden's pseudo R-squared value of 0.02 for our model, we conclude that our model has weak predictive capability.

Key Takeaways

1. The ordinal regression model to predict physical attack severity levels was valid and met all assumptions. All of the model coefficients were statistically significant. However, it has weak predictive capability, based on its pseudo R-square value.
2. According to the model output, adolescents who are male, have indicated that other students are not kind and helpful, have parents who do not understand their problems, or have a low number of close friends, are more likely to be physically attacked in bullying.

Question 3: Classification Analysis: Can we predict being bullied by pattern recognition based on various predictors?

In Q1 and Q2, we found factors associated with the severity of physical attacks in bullying and factors associated with an adolescent being bullied. Regression analysis aims to model and understand the relationship between the response variable and predictors. On the other hand, classification analysis focuses on predicting the class in which the observation belongs to. In this case, we are interested in predicting whether each student was bullied or not, based on patterns or characteristics observed in the data. The goal is to develop a classification model that not only can accurately predict whether a student was bullied or not based on our dataset, but also for novel datasets.

Data Processing and Methods

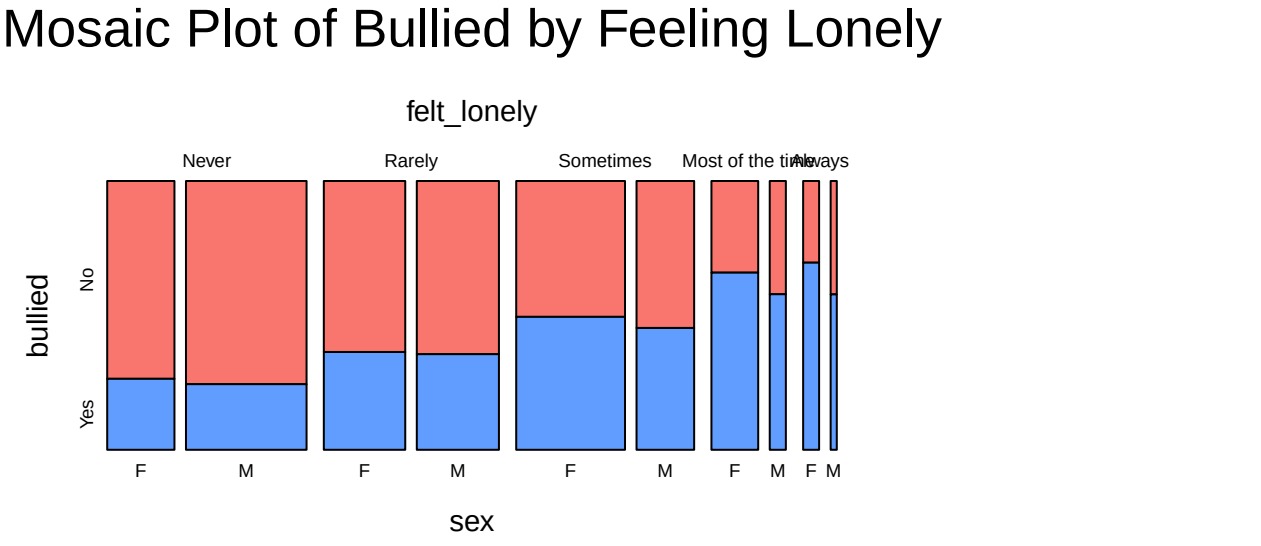
Using the wrangled data set from the ordinal logistic regression in question 2, we added a new variable, `bullied`, that is essentially a variable that represents whether an individual was bullied, in any form.

There are three variables that deal with bullying: `bullied_on_school_property_in_past_12_months`, `bullied_not_on_school_property_in_past_12_months`, and `cyber_bullied_in_past_12_months`. The response variable of interest, `bullied`, is coded `Yes` if at least one of the bullying variables has a value of `Yes`. Otherwise, `bullied` is coded to be `No`. For this analysis, we are focused on the patterns observed in the data that can help correctly categorize students based on whether or not they were bullied in general or not.

Exploratory Data Analysis

In order to deduce whether it would be beneficial to perform classification analysis, we created some mosaic plots to see whether the levels in our data set's variables differ depending on bullying status.

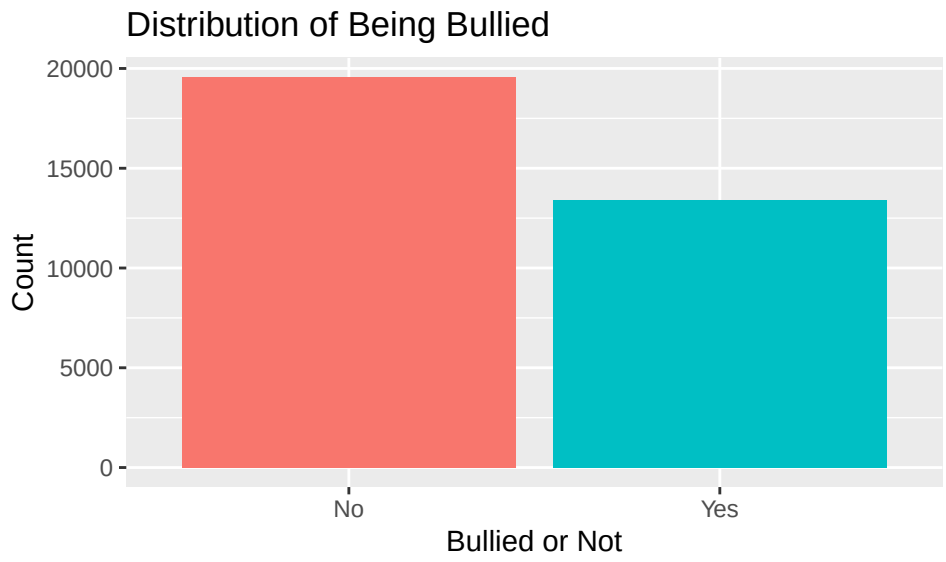
```
# mosaic plot
mosaic(~ felt_lonely + sex + bullied, data = classification_data, highlighting = "bullied",
       highlighting_fill = c("#F8766D", "#619CFF"), direction = c("v", "v", "h"),
       gp_varnames = gpar(fontsize = 11, fontface = 1.5),
       gp_labels = gpar(fontsize = 6.5), main = "Mosaic Plot of Bullied by Feeling Lonely")
```



This is a mosaic plot that displays the proportion of whether you are bullied or not based on the degree to how lonely you are (never, rarely, sometimes, most of the times, and always), faceted by sex. As the degree of feeling lonely increases (from never to always), the proportions of males and females that are bullied have a general increase in trend. However, the trend is more prominent in females.

Since we notice some differences in the proportions of students being bullied or not bullied, based on different categories of loneliness and sex, we know that there are patterns in the data that can help predict bullying status. Thus, it would be beneficial to perform classification analysis.

```
# bar chart
ggplot(data = classification_data, aes(x = bullied, fill = bullied)) +
  geom_bar() +
  labs(x = "Bullied or Not", y = "Count",title = "Distribution of Being Bullied") +
  theme(legend.position = "none")
```




```
# tally of bullied
classification_data$bullied %>%
  tally()
```

```
## X
##    No   Yes
## 19553 13385
```

```
19553/32938
```

```
## [1] 0.5936305
```

```
13385/32938
```

```
## [1] 0.4063695
```

We then created a bar chart and tally chart to investigate the distribution of the response variable for whether a student has been bullied or not. From the bar chart, we see that there are more students who

have not been bullied than students who were bullied. Looking at the tally chart, 59.4% of students have not been bullied while 40.6% of students have been bullied. Overall though, there is not a huge class imbalance (where one category is much greater than the other). Thus, our classification model will not be as biased.

```
# misclassification error rate
n <- nrow(classification_data)
(n-19553)/n
```

```
## [1] 0.4063695
```

Lastly, before creating the classification model, we calculated the misclassification rate. This is the error rate if the model classified every student as not bullied (the majority class). The misclassification rate is 40.6%, which is the baseline error rate that we want to improve upon.

Creating the Classification Tree

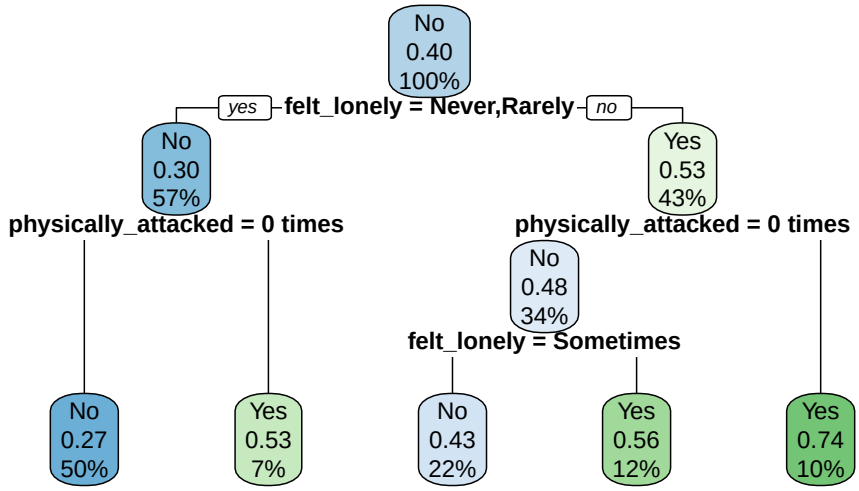
There are many different models used for classification analysis. However, we decided to use a classification tree model, since the output is visual and can be better understood by different audiences.

```
# creating training and test sets
set.seed(320)
n <- nrow(classification_data)
train_index <- sample(1:n, 0.75*n)
test_index <- setdiff(1:n, train_index)

train <- classification_data[train_index, ]
test <- classification_data[test_index, ]
```

We split 75% of the bullied data set to be a training set used to train the classification model. The training set is used to build and optimize the model's parameters and learn underlying patterns/relationships between the features and class labels. The purpose of the training set is for the model to be exposed to our data and learn the patterns and boundaries that will help make accurate predictions on unseen data.

The remaining 25% of the bullied data set is the test set, which is the unseen data. The test set is used to evaluate the performance of the trained classification model. It is an independent data set that assesses how well the classification model generalizes to unseen data.

```
# creating classification tree
bullied.tree <- rpart(bullied ~ ., data = train, method = "class")
#print(bullied.tree)
rpart.plot(bullied.tree, cex = 0.75)
```




```
printcp(bullied.tree)
```

```
##
## Classification tree:
## rpart(formula = bullied ~ ., data = train, method = "class")
##
## Variables actually used in tree construction:
## [1] felt_lonely         physically_attacked
##
## Root node error: 9960/24703 = 0.40319
##
## n= 24703
##
##         CP nsplit rel error  xerror      xstd
## 1 0.070984      0   1.00000 1.00000 0.0077408
## 2 0.040562      1   0.92902 0.92902 0.0076379
## 3 0.034940      2   0.88845 0.89819 0.0075843
## 4 0.012249      3   0.85351 0.85351 0.0074970
## 5 0.010000      4   0.84127 0.84669 0.0074826
```

In order to use the tree to make a prediction about bullying status for any adolescent, follow the guidelines:

1. For each node, if the condition is true for an adolescent, you continue down the tree by going left. If the condition is false, you continue down the tree by going right.
2. After following the tree all the way to one of the end nodes, you are left with either "Yes" or "No" for whether the adolescent is bullied or not.

- Let's run through an example with a hypothetical adolescent:
  - 1st node: `never or rarely felt lonely` = yes, so go left (←)
  - 2nd node: `never physically attacked` = yes, so go left (←)
  - 3rd node: `bullied` = No

Takeaway: An individual who never or rarely felt lonely and was never physically attacked would be predicted as **hasn't been bullied**.

- Another example:
  - 1st node: `never or rarely felt lonely` = no, so go right (→)
  - 2nd node: `never physically attacked` = yes, so go left (←)
  - 3rd node: `sometimes felt lonely` = no, so go right (→)
  - 4th node: `bullied` = Yes

Takeaway: An individual who was not physically attacked but felt lonely most of the times or always would be predicted as **has been bullied**.

Model Diagnostics

```
# apparent error rate - training set
bulliedtreepred <- predict(bullied.tree, newdata = train, type = "class")
tally(bulliedtreepred ~ bullied, data = train)
```

```
##                bullied
## bulliedtreepred    No   Yes
##             No  12035  5671
##             Yes  2708  4289
```

```
(2708+5671)/24703 # from confusion matrix
```

```
## [1] 0.3391896
```

```
0.84127*9960/24703 # from the output, should agree
```

```
## [1] 0.3391916
```

```
apparent_error <- 0.84127*9960/24703
```

The apparent error rate is 33.9%. This error rate is found from the classification model trying to predict the training set, so it is usually more generous than the true error rate.

```
# estimated true error rate - test set
bulliedtreepred2 <- predict(bullied.tree, newdata = test, type = "class")
tally(bulliedtreepred2 ~ bullied, data = test)
```

```
##                 bullied
## bulliedtreepred2   No  Yes
##              No  3984 1956
##              Yes  826 1469
```

```
(826+1956)/8235 # from confusion matrix
```

```
## [1] 0.3378264
```

```
0.85065*9960/24703 # from the output
```

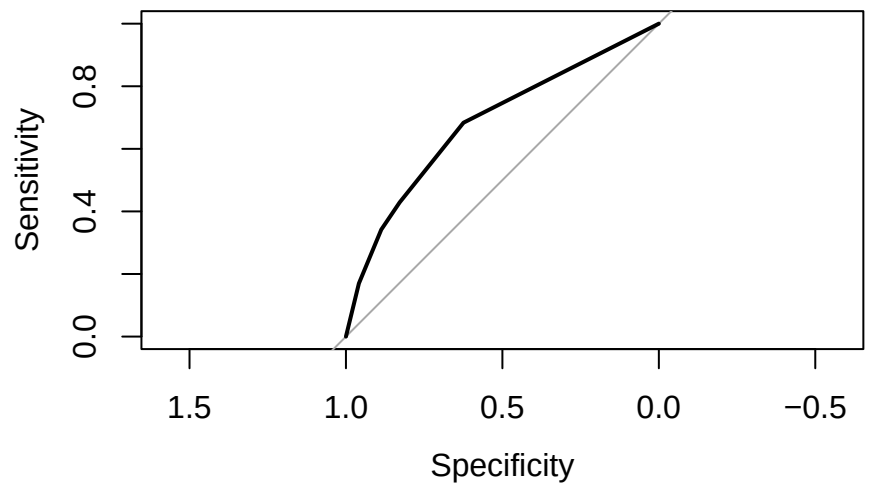
```
## [1] 0.3429735
```

```
TER <- 0.84669*9960/24703
TER
```

```
## [1] 0.3413769
```

The estimated true error rate is 34.1%. This error rate, which is found from the test set, is a 6.5% improvement from the misclassification error rate. Therefore our classification tree does a better job in predicting whether an adolescent is bullied or not bullied than if the model were just to guess that everyone belongs in the majority class.

```
# ROC and auc
predict_prob <- predict(bullied.tree, test, type = 'prob')
auc <- auc(test$bullied, predict_prob[,2])
plot(roc(test$bullied, predict_prob[,2]))
```




```
auc
```

```
## Area under the curve: 0.6835
```

An ROC curve (receiver operating characteristic curve) is a graph that shows the performance of a classification model. It shows the trade-off between the true positive rate and false positive rate as the classification threshold is varied. An ideal classifier that perfectly separates the two classes (bullied or not bullied) would have an ROC curve that passes through the top-left corner of the plot, which indicates high true positive rate and low false positive rate. The auc (area under the curve) is a measure of the ROC curve. It quantifies the model's overall performance by calculating the area under the curve. It ranges from 0 to 1, where 1 represents a perfect classifier and 0.5 corresponds to a random classifier.

The ROC curve does not exhibit as good of a curvature we would like because the curve reaches a sensitivity of 0.6-0.8 and specificity of around 0.4. The moderate curve indicates that the performance can be improved upon but not that inaccurate. The area under the curve is 0.6835, which is on the cusp of poor discrimination and acceptable discrimination.

Key Takeaways

- The predictive model we generated through a decision tree for binary response variable includes the variables relating to physical attacks and loneliness.
- There are specific levels of these categorical variables that are associated with being bullied. Any combination of being physically attacked and felt lonely most of the times or always are associated with the student being bullied.
- The estimated TER (approximated true error rate) of the tree is 34.1%. This is the proportion of mistakes made on the testing set, which is the portion of the bullied data set that the model was not made from, making it "new" data for the model performance to be calculated. This is a 6.5% improvement in error rate than if we were to just guess that all students have gotten bullied.
- The classification model has acceptable discrimination (auc ≈ 0.7), which refers to the level of accuracy or predictive performance that is considered satisfactory for this problem of bullying. This implies that our model is able to sufficiently distinguish between different classes/categories with a reasonable degree of accuracy.

Conclusion

Through our statistical analyses, we were able to answer three questions about bullying-related factors in adolescents. Based on the logistic regression models, we found that an adolescent's age, level of loneliness, frequency of missing school without permission, perception of other students as kind and helpful, level of understanding from parents, and whether or not the adolescent is overweight are significantly associated with being bullied in person. Additionally, an adolescent's level of loneliness, frequency of missing school without permission, perception of other students as kind and helpful, level of understanding from parents, and an adolescent being male are significantly associated with being cyberbullied. From our ordinal logistic regression model, we found that adolescents who are male, have indicated that other students are not kind and helpful, have parents who do not understand their problems, or have a low number of close friends, are more likely to be physically attacked in bullying. Lastly, we built a classification model with an accuracy rate of 65.8%, which can predict whether an adolescent is bullied or not using information about physical attacks and loneliness.

One limitation of our analyses is that the data was only collected from adolescents from Argentina. Thus, our models and results may not be generalizable to adolescents from other countries and cultures. Another limitation is that the logistic regression model and ordinal logistic regression model have low predictive power, based on McFadden's pseudo R-squared values.

Based on our results, we have a few recommendations for schools and parents that may prevent or reduce bullying in adolescents:

1. Schools should provide more social support and community building for adolescents to reduce loneliness and increase peer bonding. This can be in the form of peer support groups that meet at a fixed time, more structure for extracurricular activities and clubs that adolescents can join, and having student ambassadors that organize social events.
2. Parents should seek to spend more time with their children and work to understand their problems, in order to be a support system their adolescent can turn to.
3. Schools should work to foster an inclusive environment in classrooms, so all students, regardless of weight or appearance, feel safe and accepted. This can be achieved through having community guidelines in classrooms, having teachers model inclusive language and behaviors, and providing accessible mental health and counseling resources.