

# Multinomial Logistic Regression - COMP 4442 Final Project

E. Sullivan, B. Bottle, T. Mark, N. Kwartei Quartey, Peyton Capristo

2023-08-14

Install necessary packages

## Motivation

The purpose is to demonstrate the method of applying Multinomial Logistic Regression to an appropriate use case. We will walk through the setup, assumptions and their assessment, training a multinomial logistic regression model, and assessing its performance.

## Data Description

We are using data on adolescent insomnia obtained from Kaggle at the link below. This data set provides information on 95 subjects and their responses to multiple self-reporting questionnaires. Questions on the surveys include sleep quality and habits, sleep-related cognition, stress, and stress re-activity, coping and emotion regulation behavior, mood, personality, and childhood trauma. Each participant in the study completed 18 different surveys designed to assess various potential insomnia factors. Each survey was self-reported by the individual and the study involved a clinical interview for the insomnia diagnosis, which is listed as the SubGroup variable.

For our model, we decided to focus on the participants responses to the Adolescent Stress Questionnaire (ASQ) with the SubGroup diagnosis as the response variable. There are three categories in the SubGroup diagnosis: Control/No Insomnia, Sub-clinical Insomnia, and Clinical Insomnia

Data source: [https://www.kaggle.com/datasets/utkarshx27/insomnia-symptomatology-in-adolescence?select=insomnia\\_data\\_dictionary.csv](https://www.kaggle.com/datasets/utkarshx27/insomnia-symptomatology-in-adolescence?select=insomnia_data_dictionary.csv)

## Hypotheses

The hypotheses for a multinomial logistic regression model are the same as those for a binary logistic regression model

Null Hypothesis: All the coefficients for our predictor variables are equal to 0 in the population

Alternative Hypothesis: At least one of the coefficients for the predictor variables is different than 0 in the population

## Data formatting and manipulation

Load the data from insomnia\_data.csv. Modify the path as needed.

```
## 'data.frame':    95 obs. of  11 variables:
## $ SubGroup      : Factor w/ 3 levels "0","1","2": 1 1 1 1 3 1 1 1 1 1 ...
## $ Home          : int   13 15 19 12 26 12 29 12 24 13 ...
## $ School        : int   11 9 20 9 18 12 15 11 9 9 ...
## $ Attendance    : int    3 3 6 3 6 4 4 4 5 5 ...
## $ Romantic      : int    5 5 8 9 15 5 16 5 7 5 ...
## $ Peer          : int    7 7 8 9 10 8 17 7 9 7 ...
## $ Teacher       : int    7 7 10 7 8 7 16 7 7 9 ...
## $ Future        : int    6 5 6 8 7 6 9 6 8 9 ...
## $ Leisure       : int    6 5 10 5 17 10 16 6 12 11 ...
## $ Finance       : int    4 4 7 4 10 4 11 4 5 4 ...
## $ Responsibility: int    3 5 4 3 4 3 10 3 5 3 ...
```

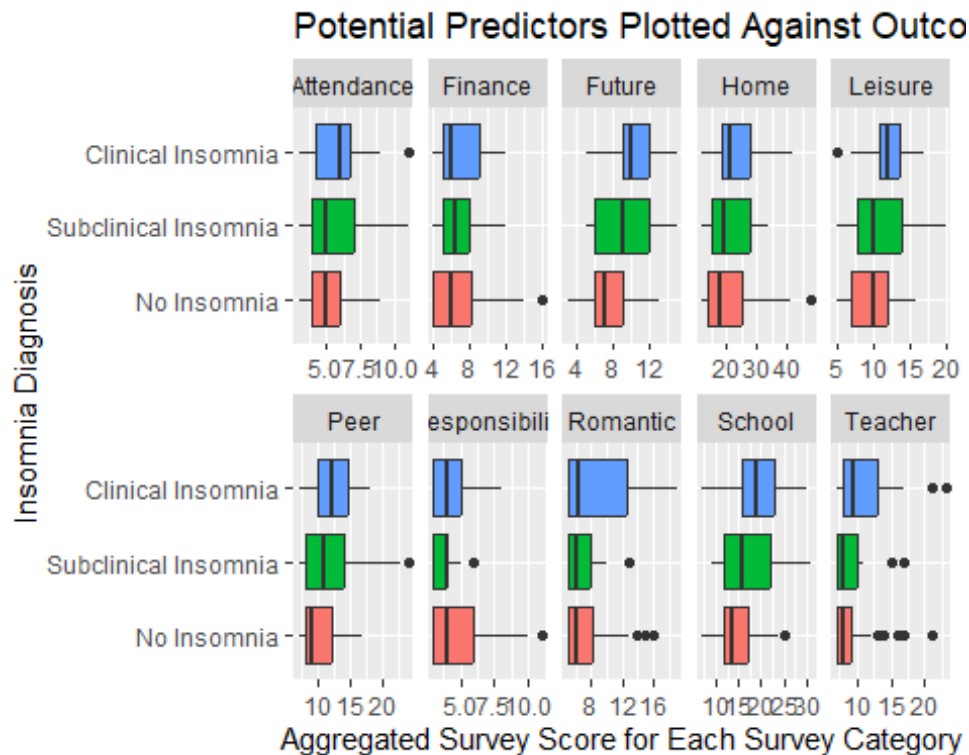
Explanation: We have selected 11 columns from the original data set: SubGroup and all columns with header that starts with the string 'asq'. Using the str() function, we can see we now have a data frame with 95 rows and 11 columns. Subgroup, our predictor variable, is set to factor-type variable with 3 levels: 0 - No Insomnia, 1 - Sub-clinical Insomnia, 2 - Clinical Insomnia.

For this demonstration, the explanatory variables are being treated as roughly continuous since we are assessing whether a point increase in the aggregated survey score for each grouping of questions correlates to an increased likelihood of an insomnia diagnosis.

## Visual review of data

As a first step, we will take a visual look at our data to see where there may be insights or concerns.

```
## Warning: Removed 2 rows containing non-finite values (`stat_boxplot()`).
```



Explanation: Box plots are a useful tool for examining categorical data since they can easily show difference between groups. Here we should pay particular attention to the median line to look for predictors that show significant differences between groups. Here it looks like school and future could potentially be useful predictors given the differences. Box plots can also alert us to potential outlier concerns. In this case, it looks like teacher could be have some problematic outliers, but we'll test that in the assumption section below.

## Additional data cleanup

One thing to note that is not strictly related to the box plot is that when we run the code, we get a warning that there are non-finite values present. This could cause problems for the model so we'll take a look at those before proceeding.

```
## SubGroup Home School Attendance Romantic Peer Teacher Future Leisure
Finance
## 42      1  34    31          9        7  24    17    14    14
NA
## 95      1  33    27         11        7  23    15    12    20
12
## Responsibility
## 42              6
## 95             NA
```

Explanation: We can use the `complete.cases()` method from stats to identify any rows with missing values. Since there are only two, and this demo is focused on the method, we'll go ahead and drop those to do a complete cases analysis. This leaves us with 93 observations.

## Assess Data Assumptions for Multinomial Logistic Regression

### Independence and Exhaustive categories

Because our observations are from a single point in time, and there is not reason to believe that there would be and dependency among the survey respondents, we will assume independence for this code demo. We also know that the study specifically looked at diagnosing No Insomnia, Sub-clinical, or Clinical Insomnia. So we know that, by study design, these categories are exhaustive.

### Minimum of 10 observations for each level and no level has more than 5 times the smallest number of observations.

Because all of our predictor variables are being treated as continuous, we simply need to verify that each level of the outcome variable has a minimum of ten observations and then compare the largest and smallest groups to ensure they are not more than 5 times different.

Note, that if we had treated any of our predictor variables as categorical we would need to ensure that each outcome/categorical pairing had a minimum of 10 observations as well.

```
##  
##  0  1  2  
## 48 19 26  
## [1] FALSE
```

Explanation: Looking at the table, we can see that our smallest group is 19 and our largest is 48. Since our smallest group is greater than 10, we've met the first part of the assumption. A simple logical test of the maximum and minimum values confirms that the largest value is not more than 5 times greater than the smallest.

### No Multicollinearity

In a situation with more than 2 predictors, a variance inflation factor assessment is the best option for assessing multicollinearity. As with all regression models, this an important assumption because the presence of multicollinearity can result in the 'bouncing betas' problem.

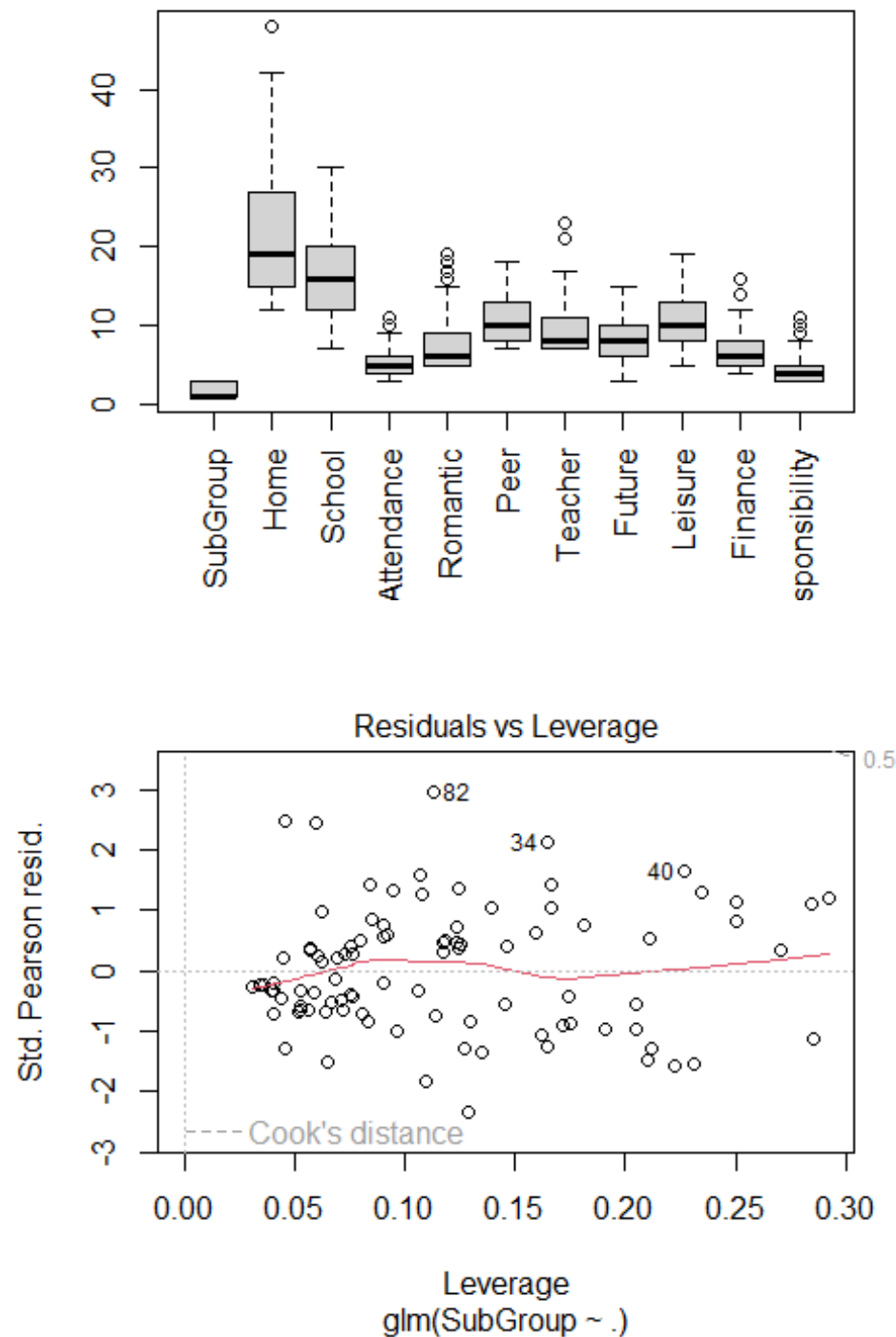
The `vif()` method does not work on models built using the `multinom` function that we will be using to build our model in the next section. However, this does not rely on the dependent/outcome variable so we can run a different regression model that works with `vif()` and use that assessment for this assumption. In this case, we've chosen to use a one-against-many binary logistic model.

##	Home	School	Attendance	Romantic	Peer
##	2.641587	2.620197	1.944094	1.637725	1.549641
##	Teacher	Future	Leisure	Finance	Responsibility
##	1.840323	2.251130	2.617700	2.179288	2.270668
##	Home	School	Attendance	Romantic	Peer
##	FALSE	FALSE	FALSE	FALSE	FALSE
##	Teacher	Future	Leisure	Finance	Responsibility
##	FALSE	FALSE	FALSE	FALSE	FALSE

Explanation: While different models and different scenarios will call for different thresholds, a general rule of thumb is that values greater than 5 are potentially concerning and greater than 10 are very concerning. By looking at both the raw values and using a simple logical statement, we can confirm that none of our values are above 5 and so we do not have any concerning levels of multicollinearity

### No Outliers, High Leverage Values, or Highly Influential Points

As we saw earlier, boxplots can be a useful way to look for potentially concerning outliers. If we see anything concerning, we can use Cook's Distance to determine if these outliers are highly influential, and therefore a concern for our model. Again, the `plot()` function that we'll use here doesn't work for models built with the `multinom` function, but a binary logistic model will still show us if there are any concerning outliers.



Explanation: In examining the box plots, we do see a few observations that appear to be outliers that could be cause for concern. so we will look at the Residual-Leverage plots to assess if these outliers are highly influential by looking for any value that crosses the dotted line for Cook's distance. Reviewing the residual-leverage plot, we do not have any

points that cross the 0.5 or 1.0 Cook's distance lines, so our data set does not violate this assumption.

## Hosmer and Lemeshow Goodness of Fit Test

Finally, as a validation we can use the Hosmer Lemeshow test to assess the validity of our binary model. While this isn't explicitly an assumption of multinomial regression, it can add a layer of confidence that if the binary model is valid our multinomial model will be as well. This test assesses whether the proportion of observations in 'g' categories matches the expected proportion of observations, if the results are too far off from expected we would get a small p-value which would indicate our model was invalid. This is most often used to measure model fit, but here just provides us with additional confidence that we've properly assessed the assumptions.

```
##  
## Hosmer and Lemeshow goodness of fit (GOF) test  
##  
## data:  vif_model$y, fitted(vif_model)  
## X-squared = 12.696, df = 8, p-value = 0.1228
```

Explanation: Our model results in a p-value of greater than .05, so we do not reject the null hypothesis and determine that our model is valid. Now that we have assessed the assumptions and provided an additional level of validation with the Hosmer Lemeshow test, we can move on to building our multinomial regression model.

## Building the multinomial model

To fit a multinomial regression model we've used the `multinom()` function from the `nnet` package. The syntax for this is very similar to the `lm()` and `glm()` functions to fit other regression models.

When running a multinomial model, it is important that you review the output of your model training to ensure the model converges. If not, you can increase the number of iterations performed in the model by changing the `maxit` argument. The default is 100 iterations, but that may not be sufficient for larger data sets or larger models.

```
## # weights:  6 (2 variable)  
## initial  value 102.170943  
## final   value 95.059254  
## converged  
  
## # weights:  36 (22 variable)  
## initial  value 102.170943  
## iter   10 value 80.256968  
## iter   20 value 70.994329  
## iter   30 value 70.061191  
## final   value 70.061176  
## converged
```

```
## Start: AIC=194.12
## SubGroup ~ 1
##
## trying + Home
## # weights: 9 (4 variable)
## initial value 102.170943
## final value 93.814353
## converged
## trying + School
## # weights: 9 (4 variable)
## initial value 102.170943
## final value 88.438434
## converged
## trying + Attendance
## # weights: 9 (4 variable)
## initial value 102.170943
## final value 91.928723
## converged
## trying + Romantic
## # weights: 9 (4 variable)
## initial value 102.170943
## iter 10 value 91.975457
## iter 10 value 91.975457
## final value 91.975457
## converged
## trying + Peer
## # weights: 9 (4 variable)
## initial value 102.170943
## final value 90.474369
## converged
## trying + Teacher
## # weights: 9 (4 variable)
## initial value 102.170943
## iter 10 value 91.553913
## iter 10 value 91.553913
## final value 91.553913
## converged
## trying + Future
## # weights: 9 (4 variable)
## initial value 102.170943
## final value 86.188517
## converged
## trying + Leisure
## # weights: 9 (4 variable)
## initial value 102.170943
## final value 91.099605
## converged
## trying + Finance
## # weights: 9 (4 variable)
## initial value 102.170943
```



```

## final value 94.812778
## converged
## trying + Responsibility
## # weights: 9 (4 variable)
## initial value 102.170943
## iter 10 value 91.235497
## iter 10 value 91.235497
## final value 91.235497
## converged
##
##           Df      AIC
## + +Future      4 180.3770
## + +School      4 184.8769
## + +Peer        4 188.9487
## + +Leisure     4 190.1992
## + +Responsibility 4 190.4710
## + +Teacher     4 191.1078
## + +Attendance  4 191.8574
## + +Romantic    4 191.9509
## <none>         2 194.1185
## + +Home        4 195.6287
## + +Finance     4 197.6256
## # weights: 9 (4 variable)
## initial value 102.170943
## final value 86.188517
## converged
##
## Step: AIC=180.38
## SubGroup ~ Future
##
## trying + Home
## # weights: 12 (6 variable)
## initial value 102.170943
## iter 10 value 83.784748
## final value 83.784690
## converged
## trying + School
## # weights: 12 (6 variable)
## initial value 102.170943
## iter 10 value 85.254813
## final value 85.254639
## converged
## trying + Attendance
## # weights: 12 (6 variable)
## initial value 102.170943
## iter 10 value 85.988852
## final value 85.988688
## converged
## trying + Romantic
## # weights: 12 (6 variable)
## initial value 102.170943

```

```

## iter 10 value 84.021530
## final value 84.021080
## converged
## trying + Peer
## # weights: 12 (6 variable)
## initial value 102.170943
## iter 10 value 84.363889
## final value 84.363881
## converged
## trying + Teacher
## # weights: 12 (6 variable)
## initial value 102.170943
## iter 10 value 83.800283
## final value 83.795536
## converged
## trying + Leisure
## # weights: 12 (6 variable)
## initial value 102.170943
## iter 10 value 84.920079
## final value 84.919989
## converged
## trying + Finance
## # weights: 12 (6 variable)
## initial value 102.170943
## iter 10 value 85.287821
## final value 85.287817
## converged
## trying + Responsibility
## # weights: 12 (6 variable)
## initial value 102.170943
## iter 10 value 79.953427
## final value 79.953327
## converged
##
##           Df      AIC
## + +Responsibility 6 171.9067
## + +Home           6 179.5694
## + +Teacher        6 179.5911
## + +Romantic       6 180.0422
## <none>            4 180.3770
## + +Peer           6 180.7278
## + +Leisure        6 181.8400
## + +School         6 182.5093
## + +Finance        6 182.5756
## + +Attendance     6 183.9774
## # weights: 12 (6 variable)
## initial value 102.170943
## iter 10 value 79.953427
## final value 79.953327
## converged
##

```

```
## Step: AIC=171.91
## SubGroup ~ Future + Responsibility
##
## trying + Home
## # weights: 15 (8 variable)
## initial value 102.170943
## iter 10 value 79.789696
## final value 79.749115
## converged
## trying + School
## # weights: 15 (8 variable)
## initial value 102.170943
## iter 10 value 79.055746
## final value 78.914761
## converged
## trying + Attendance
## # weights: 15 (8 variable)
## initial value 102.170943
## iter 10 value 79.425045
## final value 79.416301
## converged
## trying + Romantic
## # weights: 15 (8 variable)
## initial value 102.170943
## iter 10 value 77.250269
## final value 77.243047
## converged
## trying + Peer
## # weights: 15 (8 variable)
## initial value 102.170943
## iter 10 value 77.662924
## final value 77.574270
## converged
## trying + Teacher
## # weights: 15 (8 variable)
## initial value 102.170943
## iter 10 value 76.973902
## final value 76.952204
## converged
## trying + Leisure
## # weights: 15 (8 variable)
## initial value 102.170943
## iter 10 value 79.169751
## final value 79.153776
## converged
## trying + Finance
## # weights: 15 (8 variable)
## initial value 102.170943
## iter 10 value 79.264859
## final value 79.227129
```

```

## converged
##           Df      AIC
## + +Teacher    8 169.9044
## + +Romantic    8 170.4861
## + +Peer        8 171.1485
## <none>         6 171.9067
## + +School      8 173.8295
## + +Leisure     8 174.3076
## + +Finance     8 174.4543
## + +Attendance  8 174.8326
## + +Home        8 175.4982
## # weights: 15 (8 variable)
## initial value 102.170943
## iter 10 value 76.973902
## final value 76.952204
## converged
##
## Step: AIC=169.9
## SubGroup ~ Future + Responsibility + Teacher
##
## trying + Home
## # weights: 18 (10 variable)
## initial value 102.170943
## iter 10 value 80.723612
## final value 76.605089
## converged
## trying + School
## # weights: 18 (10 variable)
## initial value 102.170943
## iter 10 value 77.730203
## final value 76.760611
## converged
## trying + Attendance
## # weights: 18 (10 variable)
## initial value 102.170943
## iter 10 value 78.580533
## final value 76.545089
## converged
## trying + Romantic
## # weights: 18 (10 variable)
## initial value 102.170943
## iter 10 value 78.939011
## final value 75.167631
## converged
## trying + Peer
## # weights: 18 (10 variable)
## initial value 102.170943
## iter 10 value 77.595358
## final value 75.156867
## converged

```

```

## trying + Leisure
## # weights:  18 (10 variable)
## initial value 102.170943
## iter  10 value 80.389193
## final  value 76.672267
## converged
## trying + Finance
## # weights:  18 (10 variable)
## initial value 102.170943
## iter  10 value 77.106067
## final  value 76.145719
## converged
##           Df      AIC
## <none>      8 169.9044
## + +Peer      10 170.3137
## + +Romantic  10 170.3353
## + +Finance   10 172.2914
## + +Attendance 10 173.0902
## + +Home      10 173.2102
## + +Leisure   10 173.3445
## + +School    10 173.5212

## Call:
## multinom(formula = SubGroup ~ Future + Responsibility + Teacher,
##           data = asq_dat, maxit = 100)
##
## Coefficients:
## (Intercept)      Future Responsibility      Teacher
## 1  -0.8096247  0.3600406      -0.8745463  0.0577807
## 2  -4.1305286  0.4111647      -0.5584826  0.2591168
##
## Residual Deviance: 153.9044
## AIC: 169.9044

```

Explanation:

Since this demonstration is intended to focus on the method, we used a very minimal model selection process and chose not to use a train/test/validate split. That would be a valid approach, and generally recommended, for using this statistical method in a real world situation to help assess the model's predictive power.

Using forward selection, our final model includes future, responsibility, and teacher. Since these are the predictors we'll be using to interpret the model, we'll take a moment to define them more clearly.

Future: A rating of how much stress an adolescent experiences related to future uncertainty.

Responsibility: A rating of how much stress an adolescent experiences related to emerging adult responsibilities.

Teacher: A rating of how much stress an adolescent experiences related to teacher interaction.

## Reviewing model output

As with other regression models, our output includes coefficient estimates as well as a p-value indicating whether there is evidence to support rejecting the null hypothesis. However, with a multinomial regression model, each outcome category has its own set of coefficients and p-values associated with it.

As with binary logistic regression, the coefficient outputs of multinomial regression represent the change in log odds caused by an increase in that predictor. A positive number indicates an increase in the log odds given a one-unit change in the predictor and a negative coefficient indicates a decrease in log odds for a one-unit change in the predictor, holding all other variables constant.

```
## # A tibble: 8 × 6
##   y.level term          estimate std.error statistic  p.value
##   <chr>   <chr>          <dbl>    <dbl>    <dbl>    <dbl>
## 1 1      (Intercept)    -0.810    1.57    -0.516  0.606
## 2 1      Future         0.360    0.121     2.98  0.00290
## 3 1     Responsibility  -0.875    0.329    -2.66  0.00778
## 4 1      Teacher       0.0578   0.153     0.379  0.705
## 5 2      (Intercept)    -4.13     1.26    -3.28  0.00104
## 6 2      Future         0.411    0.115     3.57  0.000359
## 7 2     Responsibility  -0.558    0.220    -2.54  0.0110
## 8 2      Teacher       0.259    0.126     2.06  0.0392
```

Explanation: Reviewing the outcomes, and specifically the p-values, we find that there is evidence that future and responsibility have a statistically significant impact on a sub-clinical diagnosis while future, responsibility, and teacher have statistically significant impacts on a clinical diagnosis.

Therefore, interpreting the coefficients, we could say that a 1 point increase in the aggregated survey score for how much stress an adolescent experiences related to future uncertainty is correlated with a .36 increase in the log odds of an adolescent being diagnosed with sub-clinical insomnia (1) as opposed to no insomnia (0) and a .41 increase in the log odds of an adolescent being diagnosed with clinical insomnia (2) as opposed to no insomnia (0).

We could also say that a 1 point increase in the aggregated survey score for how much stress an adolescent experiences as a result of teacher interaction is correlated with a .26 increase in the log odds of that adolescent being diagnosed with clinical insomnia (2) as opposed to no insomnia (0). However, because the p-value for teacher for sub-clinical insomnia (1) is larger .05, we would not interpret the coefficient for that diagnosis.

## Converting coefficients

These coefficients can also be converted to either an odds ratio or probabilities by performing the related calculation to achieve those numbers. A simple conversion using the exponential function produces the odds ratio of the coefficients from the model. Now, a coefficient above 1 indicates that an increase in that predictor makes that outcome more likely, while a coefficient of less than 1 indicates that it decreases the likelihood.

Note that when this conversion occurs the formula is  $\exp(\hat{y}) = \exp(\beta x)$ . Because the transformation is performed both on the coefficient and the  $x$  value for the observation, the interpretation is no longer related to a one-unit increase, but the exponentiated  $x$  value. Therefore, it is more common to just interpret the results as more or less likely, rather than relating to a unit increase

```
## # A tibble: 8 × 4
##   y.level term          estimate odds_ratio
##   <chr>   <chr>          <dbl>     <dbl>
## 1 1      (Intercept)    -0.810     0.445
## 2 1      Future         0.360     1.43
## 3 1      Responsibility -0.875     0.417
## 4 1      Teacher        0.0578    1.06
## 5 2      (Intercept)    -4.13     0.0161
## 6 2      Future         0.411     1.51
## 7 2      Responsibility -0.558     0.572
## 8 2      Teacher        0.259     1.30
```

The odds ratio can then be converted using the formula  $\frac{b}{1+b}$  where  $b$  is the odds ratio. Now a value greater than 0.5 indicates more likely and a value less than 0.5 indicates less likely. Again, this formula applies to the entire  $\beta x$  and so you should be careful in interpreting the relationship of the probability to the increase in  $x$ . It is more common to convert your prediction outcomes to probability rather than just the coefficients, although the formula is the same for both.

```
## # A tibble: 8 × 5
##   y.level term          estimate odds_ratio probability
##   <chr>   <chr>          <dbl>     <dbl>     <dbl>
## 1 1      (Intercept)    -0.810     0.445     0.308
## 2 1      Future         0.360     1.43     0.589
## 3 1      Responsibility -0.875     0.417     0.294
## 4 1      Teacher        0.0578    1.06     0.514
## 5 2      (Intercept)    -4.13     0.0161    0.0158
## 6 2      Future         0.411     1.51     0.601
## 7 2      Responsibility -0.558     0.572     0.364
## 8 2      Teacher        0.259     1.30     0.564
```

## Analysis of the Model

A confusion matrix is a common way to assess the model. When there are more than two groups in a confusion matrix, each group has its own true positive, false negative, true negative, and false positive values. For example, when the true positive value is the count of SubGroup = 0 in both the actual and predicted data, the false negative value is the sum of the actual = 0, predicted = 1 and actual = 0, predicted = 2. We continue like this to find the rest of the values. Then we repeat the process to find the accuracy, precision, recall, and F1 score for the other SubGroup values (1 and 2 in this case).

```
## Actual
##  1  2  3
## 48 19 26

##      Predicted
## Actual  0  1  2
##      0 43  2  3
##      1 11  3  5
##      2  9  6 11
```

We will first calculate the confusion matrix scores for predictions of No Insomnia (0)

Next, we will calculate the confusion matrix scores for Sub-clinical Insomnia (1)

And finally we calculate the confusion matrix scores for Clinical Insomnia (2)

We can now compare these scores across categories

```
## Subgroup  Accuracy Precision    Recall    F1
## 1        0 0.7311828 0.6825397 0.8958333 0.7747748
## 2        1 0.7419355 0.2727273 0.1578947 0.2000000
## 3        2 0.7526882 0.5789474 0.4230769 0.4888889
```

Explanation: Calculating the score for each group has provided us with the model's predictive power for each category allowing us to compare where it does well and where it has challenges.

Accuracy is not the best measure when the data set is unbalanced. In this data set there is a large difference between positive and negative instances in the data. In this case, the actual values for SubGroup 0, 1, and 2 were 48, 19, and 26 respectively. Depending on your cut off for balanced vs unbalanced, we may not want to put much weight on the accuracy for each group. Based on the values above, the model was able to more accurately predict SubGroup = 2 cases, although they all seemed to be relatively similar.

A good classifier would have precision equal to 1. The precision will be 1 when the count of false positives is equal to 0. False positives are results that indicate a positive when the actual value is negative. The fewer false positives we have in our prediction (relative to the true positives) the more precise the model is for that case. The most precise prediction from our model was 0.6825 for SubGroup 0. SubGroup 0 had the most false positives out of all the groups, but since SubGroup 0 also had the most true positives by a large margin, the



precision value was the highest among the SubGroups. While Subgroup 0 and 2 are relatively close, Subgroup one appears to have poor precision compared to the others

Recall should also be 1 for a good classifier. Recall is the ratio of true positives to actual positives. In other words, how often did the model correctly predict a positive result for all actual positives. The recall value will decrease for actual positives that were predicted as negatives. The highest recall value was for SubGroup 0 at 0.8958. In this case, both Subgroup 1 and 2 had relatively poor recall compared to subgroup 1

A perfect classifier will have precision and recall equal to 1. This means that false positives and false negatives are both 0. If false positives and false negatives are both 0, that means that all actual positives were predicted positive and all actual negatives were predicted negative. F1 score is what is called a harmonic mean and is thought to be a better measure of the predictive power of a model than accuracy is. The highest F1 score for our model was for SubGroup SubGroup 0. Once again, Subgroup 2 performs better than Subgroup 1 but not as well as Subgroup 0.

In reviewing the confusion matrix, we would conclude that our model does relatively well at predicting Subgroup 0 but appears to struggle to differentiate between Subgroup 1 and 2. This indicates that while the variables we selected for our model may do well at identifying whether an adolescent has some type of sleep disorder, it is not nuanced enough to accurately predict whether that will be sub-clinical or clinical insomnia.

## Exploring Data Visually

To examine the relationship each variable has on the outcome variable prediction of SubGroup category, we can generate small data sets holding two of the variables constant while varying the third variable over the span of potential question responses. Then we can use these to visualize how the predicted probabilities change for each SubGroup as the input variable changes.

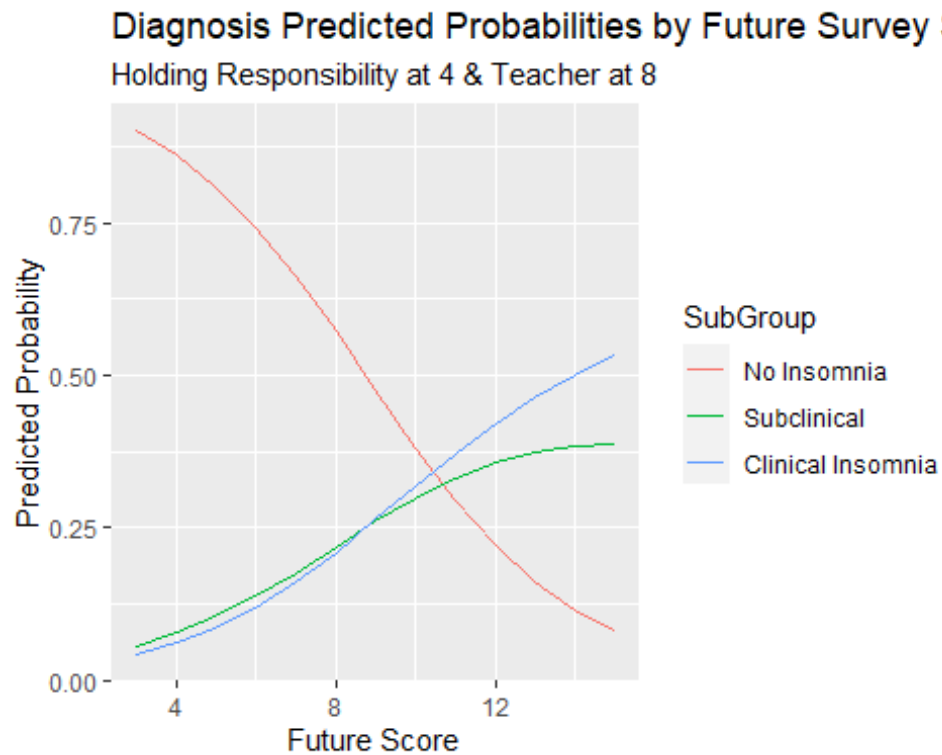
We will start by holding responsibility and teacher constant at their medians in the responses and creating a data set which contains a sequence of all possible scores for the question of future. Responsibility will be constant at 4 and Teacher will be constant at 8.

Now, holding future and teacher constant at their medians, while varying responsibility. Future and Teacher scores will be constant at 8.

Last, holding future and responsibility constant at their medians, 4 and 8 respectively, while varying teacher score over possible values.

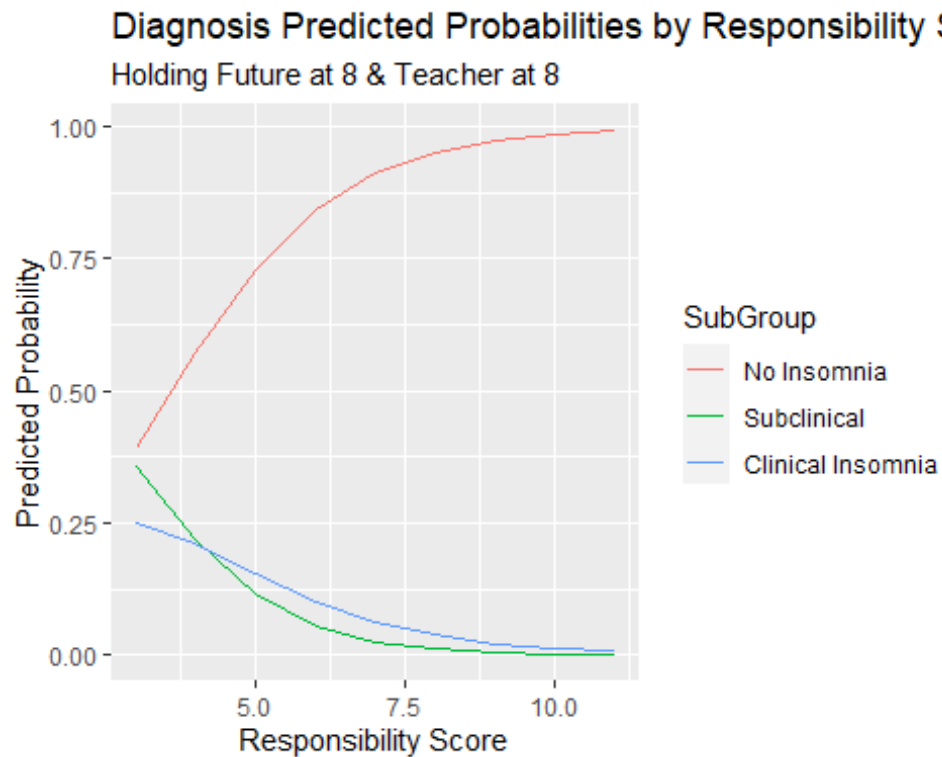
We can now plot these to see how each variable impacts the outcome predicted

## Impact of Changes in Future Score on Predicted Probability of Diagnosis



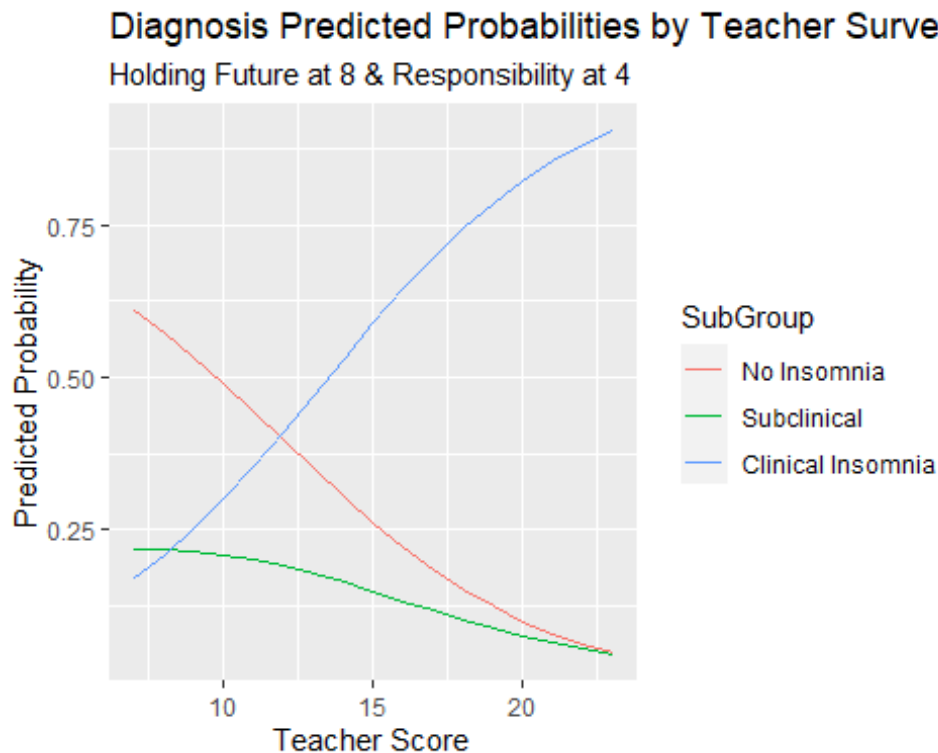
First, while holding Responsibility and Teacher survey scores constant at their medians (4 & 8 respectively), this shows how the predicted probabilities of each diagnosis (SubGroup) category changes as the survey score for Future changes. For low score responses in Future, No Insomnia has the highest predicted probability. The predicted probability of No Insomnia begins to decline until at a Future score of around 11, Clinical Insomnia becomes the predicted diagnosis. This shows that, at least for the most common Responsibility and Teacher scores, a diagnosis of Clinical Insomnia becomes more likely as the Future score increases.

## Impact of Changes in Responsibility Score on Predicted Probability of Diagnosis



With Future and Teacher scores both held constant at 8, a diagnosis of No Insomnia is the most likely outcome regardless of the Responsibility score. We can see the predicted probability of No Insomnia starts out around .4 on the low end of Responsibility and then increases quickly as the Responsibility score increase.

## Impact of Changes in Teacher Score on Predicted Probability of Diagnosis



With survey scores for Future held constant at 8 and Responsibility at 4, here we can see how changes in the Teacher survey score impact the predicted probability of each diagnosis category. While the Teacher score is low, a diagnosis of No Insomnia has the highest predicted probability. At a score of about 12, diagnoses of No Insomnia and Clinical Insomnia are predicted at nearly equal probability. However, as the Teacher score continues to increase, a diagnosis of Clinical Insomnia begins to dominate the predicted probability and becomes the most likely outcome.

### Overall

Examining these graphs as a whole, we can see a continuing theme from the confusion matrix and p-values of our model. For example, we can see that teacher has a large impact for a clinical diagnosis, but very little for a sub-clinical diagnosis, which explains the difference in p-values for the two outcomes. In addition, looking at future and responsibility we can see that sub-clinical and clinical diagnosis align very closely with very little difference between their lines. This aligns with the interpretation of the confusion matrix where we found that the model was relatively good at predicting a diagnosis of No Insomnia, but performed poorly when attempting to differentiate between sub-clinical and clinical insomnia cases.

## Conclusion

Multinomial Logistic Regression is a useful method when attempting to classify observations into 3 or more categories. It's assessment and interpretation is very similar to binary logistic regression, but because it creates multiple underlying models it is important to properly assess how the model interacts with every outcome category.