STA303 project report

Arnav Dey

2024-04-07

Research Question:Does a negative social environment have a greater impact on stress in students relative to bad physical health?

Introduction

The purpose of this project is to understand whether a negative social environment is a greater contributor to stress level in students relative to bad physical health. This question will be explored using data on student mental health with predictors measuring physiological, social, psychological, and environmental variables of students that are part of the study. Stress levels, categorized as low (levels 1 and 2) or high (level 3), will serve as the response variable. A binary logistic regression will be conducted with social predictors (peer pressure, social support, bullying, teacher-student relationships) and physiological predictors (headache, sleep quality, breathing problems). The study's significance lies in comparing how social and physiological factors influence stress perception and coping resources, aligning with the Lazarus-Folkman Transactional Stress Model. This model posits that stress results from the imbalance between perceived stressors and coping resources. The project aims to determine which set of variables has a greater influence on stress response, drawing from existing literature supporting the model's applicability(Obarius, Fischer 2021).

The results will have important implications for educational institutions like schools and universities who can develop policies to reduce student stress based on the findings of this project and many others alike.

Methods

To tackle potential problems with endogeneity, we add our psychological and environmental variables as exogenous controls. Once that is done, we fit the model with stress category as our binary response. blood pressure is not added as a predictor because stress level is calculated using blood pressure. Then, we add interaction terms for all predictors with mental health history(the only dummy variable) to see if there are any significant interaction terms at the 5% level. If there are, we add them to our model.

After having fit the model, we check for multicollinearity to ensure our predictors are not highly correlated using variance inflation factor (VIF) which measures the degree to which the variance of our slope estimates are inflated due to multicollinearity. If this number is above 5 we can consider multicollinearity to be higher than acceptable. After calculating this, we remove those variables which have a VIF above 5. We will then apply three variable selection techniques (AIC, BIC and LASSO) and obtain three different models. If any predictors of interest are removed we will add them back but if any exogenous controls are removed, we will be remove them.

We will now proceed to identify influential observations using two broad measures: DFBETAS, DFFITS and look at the number of points that fall beyond given thresholds. If the number of influential observations are less relative to the number of observations then it is not a significant problem. If there are too many influential points, we shall state that how that may skew our results.

Lastly, we will assess our 3 models using the receiver operating characteristic curve (ROC) looking at the area under the curve(AUC). If this number if close to 1 then it means our classification criterion is good. We

will then use cross-validation to analyze which model has the smallest prediction error. A final model will be picked based on its mean absolute prediction error and true positivity rate. Ideally, the former should be low and the latter should be high. We will then compare the coefficients of the final model.

Results

Numerical summaries

Sleep quality, breathing problems, and headache exhibit approximately normal distributions, with means close to medians around 2.5 on a 1 to 5 scale, suggesting moderate physical health among most students and indicating randomized, generalizable data.

Social support and bullying display centered normal distributions with slight left skew, while peer pressure and teacher-student relationship have slight right skew. Stress levels predominantly indicate low stress, with a mean of 0.99, aligning with the average scale value and the distribution of physical health data, indicating fewer high stress cases on average.

table 1: variable numerical summaries

	sleep_quality	breathing_problem	headache
	Min. :0.00	Min. :0.000	Min. :0.000
	1st Qu.:1.00	1st Qu.:2.000	1st Qu.:1.000
	Median :2.50	Median :3.000	Median :3.000
	Mean :2.66	Mean :2.754	Mean :2.508
	3rd Qu.:4.00	3rd Qu.:4.000	3rd Qu.:3.000
	Max. :5.00	Max. :5.000	Max. :5.000

social_support	peer_pressure	bullying	stress_category	teacher_student_relationship
Min. :0.000	Min. :0.000	Min. :0.000	Min. :0.0000	Min. :0.000
1st Qu.:1.000	1st Qu.:2.000	1st Qu.:1.000	1st Qu.:0.0000	1st Qu.:2.000
Median :2.000	Median :2.000	Median :3.000	Median :0.0000	Median :2.000
Mean :1.882	Mean :2.735	Mean :2.617	Mean :0.3355	Mean :2.648
3rd Qu.:3.000	3rd Qu.:4.000	3rd Qu.:4.000	3rd Qu.:1.0000	3rd Qu.:4.000
Max. :3.000	Max. :5.000	Max. :5.000	Max. :1.0000	Max. :5.000

Interaction term modeling with mental health history revealed two significant variables at the 5% level: bullying and sleep quality (one social and one physiological). Though other interactions were significant, we retained only those relevant to our research question for model simplicity. The full model is outlined below.

Table 3: Table 2: Summary of complete model

term	estimate	std.error	statistic	p.value
(Intercept)	-1.2696132	1.1371372	-1.1164996	0.2642083
anxiety_level	-0.0431091	0.0292552	-1.4735539	0.1406017
self_esteem	-0.0957427	0.0201030	-4.7626079	0.0000019
mental_health_history	-0.0757539	0.8057570	-0.0940158	0.9250966
depression	0.0463786	0.0221903	2.0900353	0.0366146
headache	0.3486858	0.1062266	3.2824725	0.0010290
sleep_quality	-0.0490441	0.1420649	-0.3452233	0.7299265
breathing_problem	0.0246633	0.0995531	0.2477405	0.8043352
noise_level	0.3913893	0.1050964	3.7240983	0.0001960

term	estimate	std.error	statistic	p.value
living_conditions	-0.0271660	0.1120856	-0.2423687	0.8084945
safety	-0.1267607	0.1180354	-1.0739211	0.2828580
basic_needs	-0.0938238	0.1089431	-0.8612187	0.3891176
academic_performance	-0.2636654	0.1075000	-2.4527011	0.0141788
study_load	0.1091976	0.1039632	1.0503489	0.2935577
teacher_student_relationship	0.2902619	0.1361362	2.1321437	0.0329950
future_career_concerns	0.0680761	0.1084764	0.6275655	0.5302886
social_support	-1.3177588	0.2037491	-6.4675578	0.0000000
peer_pressure	0.1064197	0.1117118	0.9526270	0.3407791
extracurricular_activities	0.2736305	0.1053862	2.5964547	0.0094191
bullying	-0.1248028	0.1467802	-0.8502701	0.3951750
mental_health_history:bullying	0.6602494	0.2055197	3.2125843	0.0013155
mental_health_history:sleep_quality	-0.4996752	0.1998665	-2.5000445	0.0124178

Breathing problem(a predictor of interest) is also not significant at the 5% level but will be considered anyways due to its contextual relevance.

When testing for multicollinearity with VIF, all variables had values below five except for mental health history and its interaction terms, which had VIF values of 7.37 and 7.06, likely due to inherent multicollinearity from interaction terms, inflating the mental health history coefficient. However, this is not expected to impact our model significantly. In our investigation of influential observations, 116 points exceeded the DFBETAS threshold and 149 points exceeded the DFFITS threshold, representing a small proportion (less than 15%) of the total 1100 observations, thus unlikely to bias our slope or fitted value significantly.

During variable selection with LASSO, all predictor coefficients shrank to zero, rendering the model empty. Consequently, we set aside the LASSO model for further consideration

When we apply stepwise selection using AIC, we get the following reduced model:

Table 4: Table 3: Model summary for AIC model

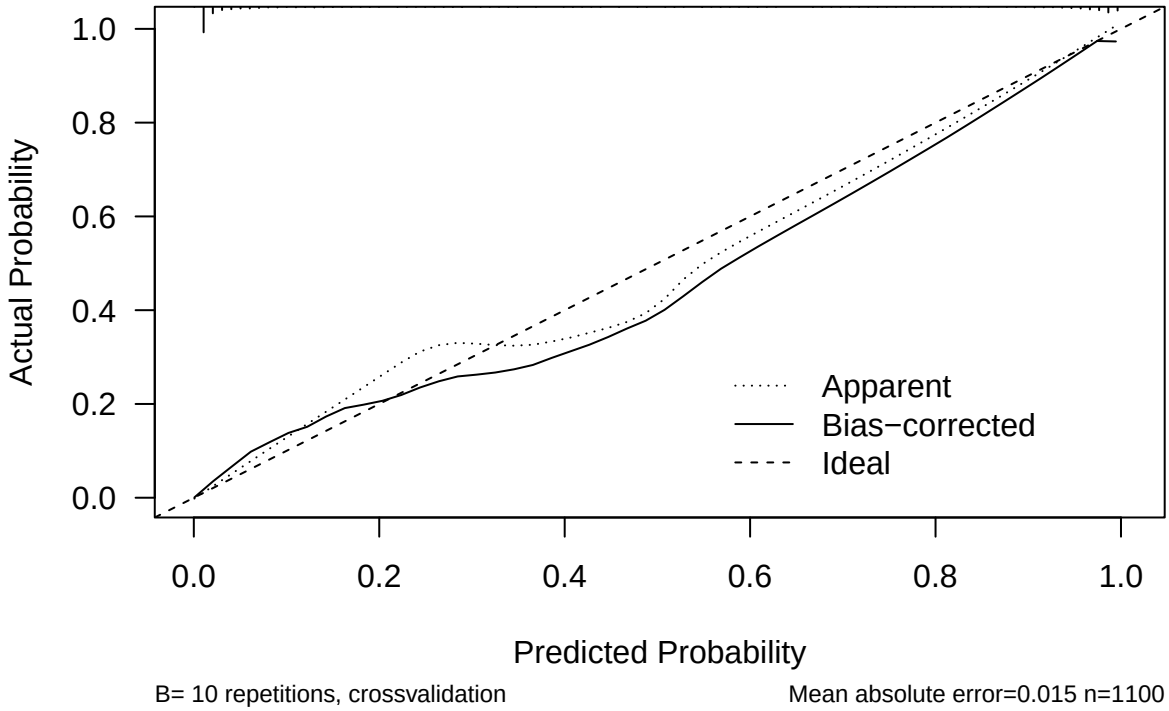
term	estimate	std.error	statistic	p.value
(Intercept)	-1.7169587	0.9233034	-1.8595824	0.0629446
self_esteem	-0.1023073	0.0199284	-5.1337369	0.0000003
sleep_quality	-0.0338136	0.1384744	-0.2441867	0.8070862
headache	0.4029232	0.1015928	3.9660590	0.0000731
noise_level	0.4072949	0.1019897	3.9934899	0.0000651
academic_performance	-0.2645828	0.1078472	-2.4533115	0.0141548
social_support	-1.2918177	0.1933224	-6.6821930	0.0000000
extracurricular_activities	0.2651271	0.1019347	2.6009509	0.0092966
bullying	-0.0975338	0.1438027	-0.6782476	0.4976147
mental_health_history	-0.1245589	0.7970582	-0.1562732	0.8758177
teacher_student_relationship	0.2625740	0.1328213	1.9768965	0.0480533
depression	0.0439499	0.0216144	2.0333680	0.0420154
breathing_problem	0.0301172	0.0967910	0.3111575	0.7556809
bullying:mental_health_history	0.7159506	0.1945400	3.6802237	0.0002330
sleep_quality:mental_health_history	-0.5399104	0.1966483	-2.7455632	0.0060407

Peer pressure is removed from the model. There are more predictors as AIC is a less harsh penalty. We also obtain the model after running stepwise selection in BIC.

After the two models were obtained we constructed the ROC curves and obtained the AUC for both. They are as follows: BIC model: 0.982 AIC model: 0.984

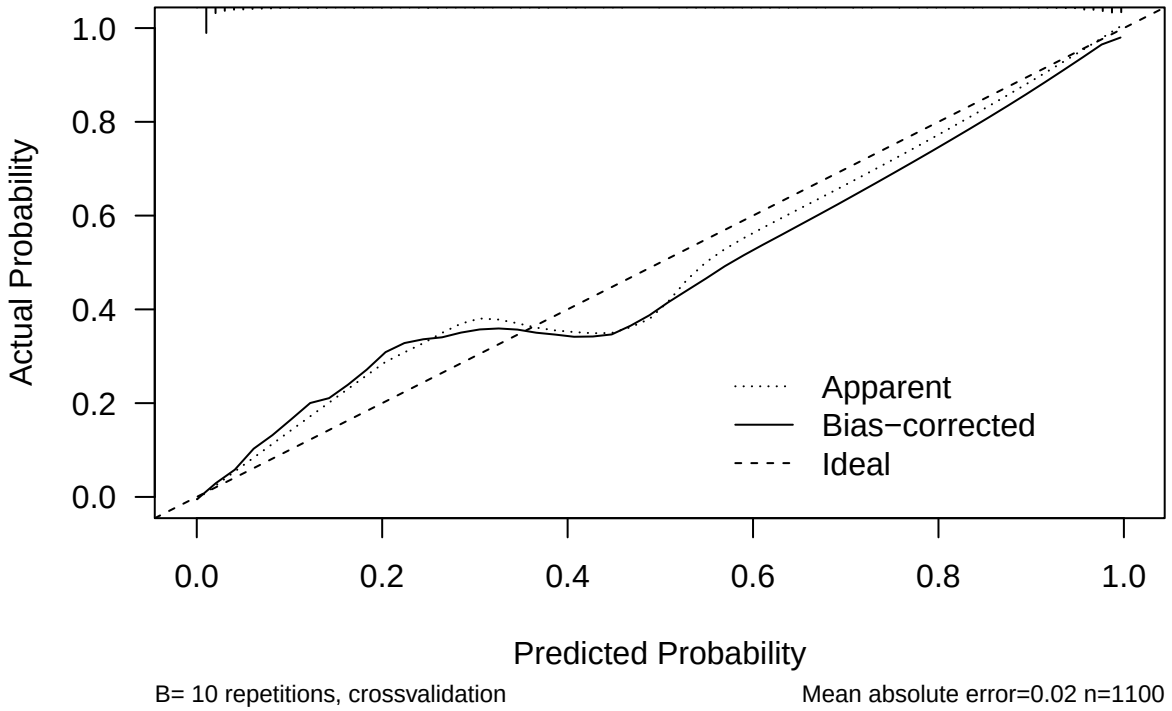
Finally, we used cross validation on both models which gave us the following results:

Graph 1: BIC model cross validation




```
##
## n=1100   Mean absolute error=0.015   Mean squared error=0.00067
## 0.9 Quantile of absolute error=0.037
```

Graph 2: AIC model cross validation




```
##
## n=1100   Mean absolute error=0.02   Mean squared error=0.00094
## 0.9 Quantile of absolute error=0.06
```

Based on model diagnostics, the AIC model slightly outperforms the BIC model with an AUC of 0.94 versus 0.92. A higher AUC indicates a better true positivity rate, especially notable since the AIC model includes more predictors. However, in cross-validation calibration plots, the BIC model exhibits a lower mean absolute error (0.019) compared to 0.022 for the AIC model.

Now, we face a tradeoff between mean absolute error and true positivity rate, crucial for identifying high-stress cases central to our research. Given the significance of accurately predicting high stress, I prioritize true positivity rate over overall prediction error, favoring the AIC model for its higher AUC/true positive rate.

Discussion

In our final model, the following is how we will calculate the odds of high stress.

A good student-teacher relationship increases the odss of stress by 1.3 times. This is done by taking the eulers constant as the base and the coefficient 0.2657 as the power. The odds increasing may be due to endogeneity issues wherein teacher-student relations do not directly affect stress, instead being governed by a different variable.

Physiological factors like sleep quality and headaches increase the odds of high stress by 1.44 and 1.496 times, respectively, for students with a mental health history. Breathing problems have a smaller impact, increasing odds by 1.03 times. In contrast, bullying significantly raises stress odds by 1.85, while higher social support lowers odds by 0.27. Conversely, a decrease in social support boosts odds by 3.7 times. On average, social variables elevate stress odds more than physiological ones, with average odds of 2.28 versus 1.32, respectively. This suggests a negative social environment contributes more to stress than health issues. With reference to the literature, social problems like bullying and low social support outweigh health problems as stressors,

with social support also acting as a crucial coping resource. Overall, social variables play a greater role in triggering stress responses.

Limitations

One potential issue is endogeneity, as stress stems from numerous variables not all considered in this model. Considering that LASSO shrunk all predictor slopes to zero suggests that there are other variables with predictive power that are not considered.

Moreover, the data is survey data which asks the student to rate their anxiety level, sleep quality etc using an artificially constructed scale. This is a somewhat arbitrary mode of measurement which makes it tougher to understand what differences from one level to another signifies.

Additionally, although the number of influential points are small relative to the sample size, the degree of influence each point has relative to each other has not been analyzed. Thus, the impact of this on parameter bias has not been assessed due to the number of influential points. word count: 1,265

Bibliography

Obbarius, Nina, et al. "A Modified Version of the Transactional Stress Concept According to Lazarus and Folkman Was Confirmed in a Psychosomatic Inpatient Sample." Frontiers in Psychology, vol. 12, 2021.

Appendix

Table 5: VIF values for each predictor

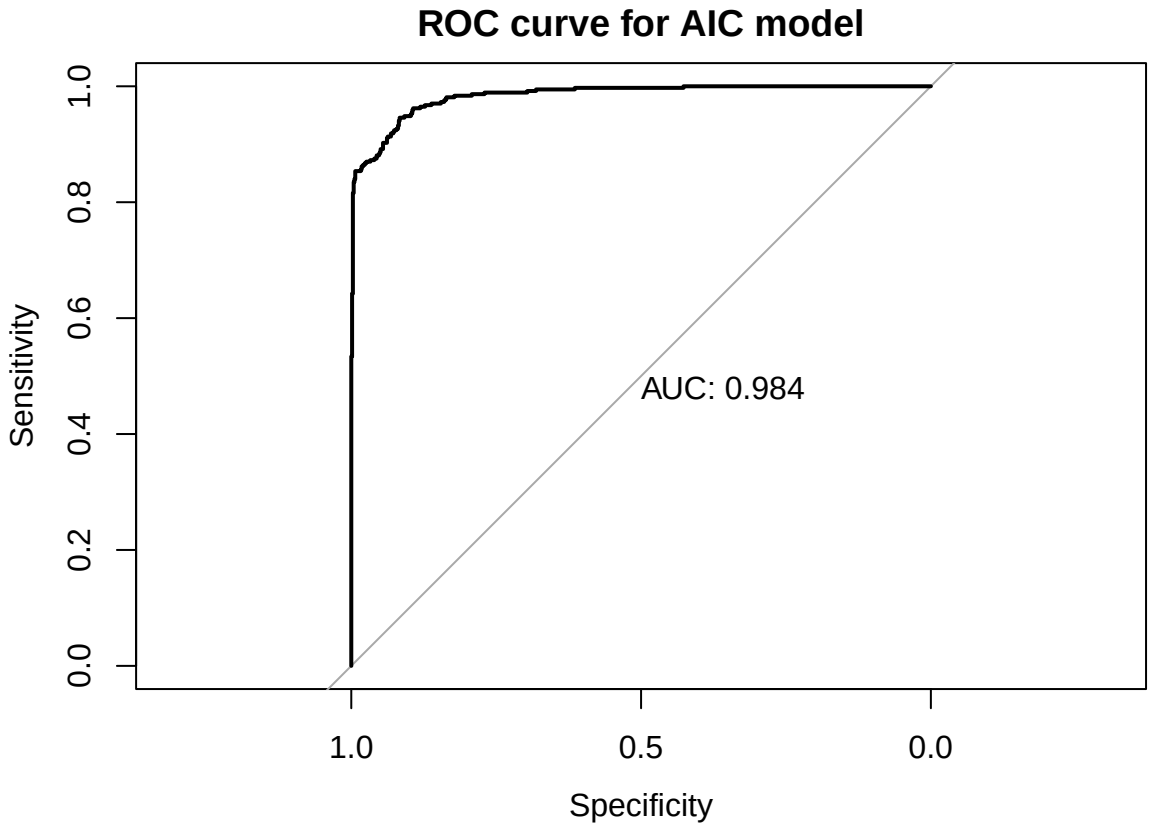
	x
anxiety_level	1.412920
self_esteem	1.171611
mental_health_history	7.374141
depression	1.290217
headache	1.170350
sleep_quality	2.310779
breathing_problem	1.227280
noise_level	1.130411
living_conditions	1.162896
safety	1.143646
basic_needs	1.208172
academic_performance	1.081786
study_load	1.169557
teacher_student_relationship	1.230422
future_career_concerns	1.285599
social_support	1.187810
peer_pressure	1.451974
extracurricular_activities	1.212791
bullying	2.559615
mental_health_history:bullying	7.059863
mental_health_history:sleep_quality	4.018135

Table 6: Model summary of BIC model

term	estimate	std.error	statistic	p.value
(Intercept)	-1.3316845	0.7548239	-1.7642321	0.0776929
self_esteem	-0.1064917	0.0192802	-5.5233586	0.0000000
sleep_quality	-0.0458583	0.1352882	-0.3389675	0.7346342
headache	0.4464845	0.1000898	4.4608387	0.0000082
noise_level	0.3747382	0.0991629	3.7790141	0.0001575
mental_health_history	-0.1174973	0.7992782	-0.1470042	0.8831287
social_support	-1.2369800	0.1877977	-6.5867680	0.0000000
extracurricular_activities	0.3028555	0.0994195	3.0462380	0.0023172
bullying	-0.1058902	0.1420813	-0.7452793	0.4561029
breathing_problem	0.0342434	0.0965949	0.3545052	0.7229603
mental_health_history:bullying	0.7731529	0.1927429	4.0113168	0.0000604
sleep_quality:mental_health_history	-0.5771433	0.1950565	-2.9588523	0.0030879

```
## Setting levels: control = 0, case = 1
```

```
## Setting direction: controls < cases
```




```
##
## Call:
## roc.default(response = df_stress$stress_category, predictor = model_AIC$fitted.values,     plot = TRU
##
## Data: model_AIC$fitted.values in 731 controls (df_stress$stress_category 0) < 369 cases (df_stress$s
## Area under the curve: 0.9837
```