ANALYSING STUDENT DEPRESSION RATES

Bayesian Inference (ST308)

Candidate Number: 51126

Introduction

The dataset I have chosen to analyse for this project is the Healthy Minds Study[1] of 2023-2024, which is created by a web-based survey given to 197 participating colleges, mainly in the USA. The study is focused on mental health issues in students, with both undergraduate and graduate students included in the sample.

The outcome I have chosen to focus on is major depression. Whether a student classifies as having major depression is decided upon by the Patient Health Questionnaire-9, where a respondent answers 9 questions related to depression (Kroenke, Spitzer, & Williams, 2001), with a scale of 0-3 for each question. A person is classed as having major depression if the sum of their scores is greater than 15 (the maximum score is 27).

Previous literature finds a number of variables significant. For example, race, sexuality, gender, age, financial struggles, living arrangement, as well as relationship status were found significant predictors of poor mental health (including anxiety) in a 2007 study of a single college believed to be representative of the USA. (Eisenberg, Gollust, Golberstein, & Hefner, 2007). My analysis similarly finds economic factors such as financial struggles and living arrangement to be very significant, along with a student's sense of belonging and whether they feel lonely. A student's identity was not found to be as significant; age did not predict any differences in depression rates, and the differences for gender and racial groups were also not significant (though sexuality was).

One limitation of the 2007 study was that although the college chosen was representative of the US student population in terms of gender and racial make-up, it was also a highly competitive, research-oriented institution, which may have swayed results. The hierarchical analysis conducted in this report clearly tackles this issue, as it models the differences between colleges, whilst also supporting generalisability of findings.

Exploratory Data Analysis

Certain columns from the dataset were more specific descriptors of other columns. For example, there were columns specifying whether Asian students were Indian, Bangladeshi, Chinese etc. Whilst these differences would have been interesting to evaluate, in the interest of simplicity, I included just the broadest descriptors of most categories to avoid potential issues with multicollinearity. I also combined certain columns, such as those representing students who identified as non-binary, having a 'queer gender' identity, and being transgender to a single column that represented students of minority gender identities. Finally, I removed students with age greater than 60, as their results may have disproportionately swayed the importance of the age predictor.[2]

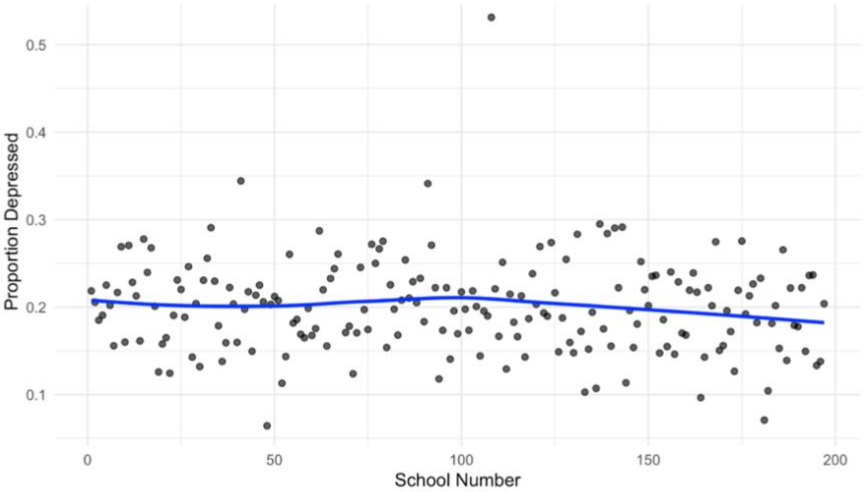
I also decided to explore the relationship between the college attended and major depression rates. Figure 1 supports the idea that there *are* disparities in depression rates across different colleges; however, perhaps not so much to warrant treating each college as a separate level. The hierarchical approach to modelling the problem presents a good middle ground; allowing

[1] https://healthymindsnetwork.org/hms/

[2] One respondent was also recorded as being 120 years old, so I suspect there may have been errors for the higher ages recorded

for an analysis of the differences between colleges, whilst also supporting generalisability to an unseen college (which is any college we might think of, since no specific colleges are specified in this dataset, for privacy reasons).

Figure 1: Major Depression rate by school number



Analysis

I created both a single level and multi-level/hierarchical model for the problem. Each model utilised a Bayesian framework, which also necessitated a discussion of the prior distributions chosen behind each model.

Single Level Model

I first modelled the problem without a hierarchical approach. That is, the model was a separate/fixed effects approach compared to the later hierarchical model; each college was treated as a separate indicator variable, independent of all other colleges.

I opted to not use any previous research to inform my prior choices for this (or the multi-level) model. This was since I had almost one hundred thousand observations and wanted my results to be informed by the data provided, rather than possibly outdated research. However, I still decided to use weakly informative priors, as flat priors could be biased to produce overly large coefficients[3] (Gabry & Goodrich, 2020). I evaluated two sets of priors; the default priors provided in rstanarm, as well as lasso priors, that would more heavily penalise possibly irrelevant predictors. After evaluating both sets of priors based on an approximation[4] of their leave-one-out (loo) cross validation accuracy (PSIS-LOO to be exact), I found the model with the lasso priors to have more predictive power than the default priors, so chose to use it as my prior going forward (the intercept prior was kept at the default). The lasso prior was a LaPlace distribution with a mean of zero, and a variance scaled based on the individual variances of different predictors.

[3] Additionally, trying to test flat priors on even a reduced size dataset in R led to an incredibly slow runtime.

[4] Although I attempted to use actual cross-validation, fitting my model even once took a long time (3 hours), which meant fitting a model multiple times was infeasible

A subset of the results from the single level model (with lasso priors) is provided below:

	mean	sd	2.5%	50%	97.5%
(Intercept)	-2.8	0.2	-3.2	-2.8	-2.4
age	0	0	0	0	0
gender_male1	-0.1	0	-0.2	-0.1	0
gender_female1	0	0	-0.1	0	0.1
gender_other1	0.4	0	0.3	0.4	0.5
sexual_h1	-0.3	0	-0.4	-0.3	-0.3
sexual_l1	0.2	0.1	0.1	0.2	0.4
sexual_g1	-0.1	0.1	-0.2	-0.1	0.1
sexual_bi1	0.2	0	0.1	0.2	0.2
sexual_queer1	0.1	0	0	0.1	0.2
sexual_quest1	0.1	0.1	0	0.1	0.3
sexual_asexual1	0.3	0.1	0.2	0.3	0.4
sexual_pan1	0.3	0.1	0.2	0.3	0.4
sexual_prefnoresp1	-0.2	0.1	-0.3	-0.2	0
sexual_selfID1	-0.1	0.1	-0.2	0	0.1
race_black1	0	0	-0.1	0	0
race_ainaan1	0.1	0.1	0	0.1	0.2
race_asian1	0.1	0	0	0.1	0.2
race_his1	0	0	-0.1	0	0
race_pi1	0	0.1	-0.2	0	0.2
race_mides1	0.2	0.1	0	0.2	0.3
race_white1	0	0	-0.1	0	0
race_other1	-0.1	0.1	-0.3	-0.1	0
international1	0.1	0.1	0	0.1	0.2
housing_worry1	-0.1	0.1	-0.3	-0.1	0

These have been truncated, as including the whole table would take up too much space. (There are more than 200 total predictors).

Surprisingly, age was not found to be a significant predictor here, contrary to previous research. The other predictors that the model found to not affect depression rate (in any direction) were alcohol consumption and year of study (so first years were predicted no different rates of depression to third years).

The most significant predictors by far were whether a student felt lonely, their grades, as well as their financial situation. For example, a particular student[5] that is predicted a 5% chance of major depression by the model and reports not feeling lonely has his chances of depression

[5] Who is a black, heterosexual male full-time student attending school100 that is not too worried about his financial situation

rise to 18% if he reports feeling lonely![6] The effect of social circumstances on student mental health was consistently emphasised in the model, with the predictor 'belonging' (the result of a student reporting whether they felt they belonged in the campus community), predicting a higher chance of major depression if a student 'majorly disagreed' with the statement that they saw themselves as part of the campus community. Surprisingly, the predictor (activ_none), which was 0 if a student attended any societies and 1 if they did not, did not predict any different rates of depression, perhaps suggesting simply encouraging more societies is not an easy fix.

Additionally, students who achieve grade As are predicted a lower chance of major depression than students with lower grades (as the grades get worse, the chance of depression rises), with students who report being grade F students having the highest chance of major depression. The effect of the specific degree chosen was not too strong; however, it seems students pursuing their MD had the highest chance of major depression (compared to bachelors, masters etc.) and students in the humanities or arts departments also had a higher rate of major depression than students studying natural sciences.

A student's economic circumstances were also significant. If students report their current financial situation to be very stressful, it predicts a higher chance of major depression (if they report their past financial situation to be stressful, it also predicts a higher chance of depression, though not nearly as much). This also holds for related factors to financial worry, such as housing insecurity. One surprising finding, that also aligned with past research, was that students who report themselves as feeling *much wealthier*[7] than the other students in their college are predicted a higher chance of major depression (coefficient of 0.3 in the model) than students who report themselves as *much poorer* than the rest of their college (coefficient of -0.1). This only held for the extreme ends of the scale though, with students who felt significantly wealthier but not the wealthiest predicted the lowest rates of depression by my model (-2.8, a *very* high coefficient for the logistic model).

A student's identity was somewhat significant, though not on the same scale as the previous factors. For example, as seen in the results table for this section, students that identified as women or 'other' were predicted higher rates of major depression than men. Additionally, lesbian, bisexual, or asexual students were predicted higher chances of depression than gay, or heterosexual students (which may have a correlation with gender). Asian, middle eastern and native American students were predicted higher rates of depression than white or black students, though not significantly so. Finally, a surprising predictor of major depression was whether students reported that they needed to be thin in order to feel good about themselves, which had a much higher coefficient than any other 'identity' predictor.

The school a student attended was also significant, with some schools predicting higher rates of depression (such as school 108), whilst others predicted lower rates (such as school 112). However, these results are not of interest to researchers, as the dataset does not provide access to descriptors of specific schools. Additionally, some schools have much less data than others, which means the specific coefficients for these schools have likely been overfit. As

[6] To be exact, a student is classed as being lonely if they answer affirmatively to multiple questions about loneliness, such as 'lacking companionship' and 'feeling isolated from others'

[7] On a scale of 1-10 students are asked to report how much wealthier they are then the rest of their cohort- I am discussing students who report being a 10 vs a 1

such, I take advantage of the partial pooling feature of hierarchical models, which avoids overfitting by shrinking school wide estimates to the global mean.

Multilevel Model

The first multilevel model I attempted to fit failed to converge, which I believed was due to a perfect separation issue, where a lack of data for predictors in certain schools led to infinite coefficients. For example, there were very few ‘gender other’ students in some schools, which meant that if the single student that identified as being ‘gender other’ was also depressed, the coefficient for being a different gender would be infinite. As a result, I decided to drastically reduce the number of predictors to the few that showed consistent variation across all schools; age, gender_male, sexual_h, race_white, activ_none (whether the student is active in any societies), lonely, and fincur (a student’s current financial situation). After this, I was able to fit a random intercept model, where the intercept of each school was different from one and other, but also taken as coming from a single distribution.

Once again I tested the default priors against the more punishing lasso priors. This time I found (when evaluated on approximate LOO cross validation error) that once again the lasso priors (LaPlace distribution with mean 0 and variance auto scaled based on predictor variance) yielded superior results, so used them again.

The model then yielded the following results:

	mean	sd	2.5%	50%	97.5%
(Intercept)	-1.4	0.1	-1.7	-1.4	-1.1
age	0	0	0	0	0
gender_male1	-0.2	0	-0.2	-0.2	-0.1
sexual_h1	-0.7	0	-0.7	-0.7	-0.6
race_white1	0	0	0	0	0.1
activ_none1	0.2	0	0.1	0.2	0.2
lonely1	1.7	0	1.7	1.7	1.8
fincur1	0.7	0.1	0.4	0.7	0.9
fincur2	-0.1	0.1	-0.3	-0.1	0.2
fincur3	-0.7	0.1	-0.9	-0.7	-0.4
fincur4	-1	0.1	-1.3	-1	-0.8
fincur5	-0.9	0.1	-1.2	-0.9	-0.6
b[(Intercept) schoolnum:2]	0	0.1	-0.2	0	0.2
b[(Intercept) schoolnum:3]	-0.1	0.1	-0.3	-0.1	0.1
b[(Intercept) schoolnum:4]	-0.2	0.1	-0.4	-0.2	0
b[(Intercept) schoolnum:5]	0.1	0.1	-0.1	0.1	0.2
b[(Intercept) schoolnum:8]	0.1	0.1	-0.2	0.1	0.4
b[(Intercept) schoolnum:10]	-0.1	0.1	-0.4	-0.1	0.2
b[(Intercept) schoolnum:11]	0.3	0.1	0.1	0.3	0.5
b[(Intercept) schoolnum:12]	0.1	0.1	-0.1	0.1	0.3
b[(Intercept) schoolnum:14]	-0.2	0.1	-0.4	-0.2	0

b[(Intercept) schoolnum:15]	0.2	0.1	0	0.2	0.4
b[(Intercept) schoolnum:16]	0	0.1	-0.1	0	0.2
b[(Intercept) schoolnum:17]	0	0.1	-0.1	0	0.2
b[(Intercept) schoolnum:18]	0	0.1	-0.2	0	0.2

Which have once again been truncated for space constraints.

Of course, these differ from the single level model by including fewer predictors, but the results are also stronger statistically speaking since they acknowledge that students in the same school are likely to have correlated results.

The model once again finds race, gender and age to only be mildly significant at the most, whilst emphasising the increased rates of major depression amongst students of minority sexualities (compared to heterosexual students). Additionally, identifying as lonely is once again very significant and the coefficient for being active in no societies is now positive (it was not in the fixed effects model), though still not a very large effect. Finally, financial situation is again significant; with students that describe their financial situation as 'always stressful' predicted higher rates of depression than any other type of student. Though the hierarchical model did not find anything majorly different or contrary to the fixed effects model, it reinforces the model's key findings.

Conclusion and Possible Extensions

Both models I built emphasised the same key findings; identity is secondary to a student's financial situation and whether they feel lonely when predicting major depression. The lack of importance of age, gender and race compared to past research (the 2007 paper) may demonstrate how their importance has changed over time; though the inclusion of a large dataset with multiple different schools could also suggest these results are more statistically robust than past research. Colleges wishing to reduce depression amongst students could therefore try providing more financial support and creating more social spaces, though this might not be as simple as encouraging more students to join societies.

One possible extension to the modelling would be to use a 'random slopes' approach to multi-level modelling, where (for example) the effect of age is dependent on college attended. Although I wanted to do this, the computational constraints of fitting such a model did not allow me to, though the results of such an approach could be very interesting.

Bibliography

Eisenberg, D., Gollust, S., Golberstein, E., & Hefner, J. (2007). Prevalence and Correlates of Depression, Anxiety, and Suicidality Among University Students. *American Journal of Orthopsychiatry*, 534-542.

Gabry, J., & Goodrich, B. (2020, July 20). *Prior Distributions for rstanarm Models*. Retrieved from https://mc-stan.org/rstanarm/articles/priors.html

Kroenke, K., Spitzer, R. L., & Williams, J. B. (2001). Patient Health Questionnaire-9 (PHQ-9).

The dataset is provided at https://github.com/codingfishguy/ST308-Coursework in a zip file.