Estimating R Anxiety Level Distribution Among Students At the University Using MRP*

Yiqu Ding

2020-12-16

Abstract

In this report, I look at different factors that affect a student's anxiety level towards using R. The data is obtained from online survey responses, which contained variables that could influence anxiety levels. I then run a multilevel regression on the sample and post-stratify them using a simulated student census. After the post-stratification, we see the estimated mean anxiety scores are higher than average anxiety scores from the raw data.

Keywords: MRP, R, Psychology, Education

Introduction

Relatively speaking, the science of statistics is a new discipline. In 1998, the public image of statistics was poor, and almost nobody knows what statisticians do(Nelder 1999). Now, statistics is an essential tool for nearly all millennial industries. Accompanied by technological improvements in computers, R has become a necessary tool for all statistical practitioners; that makes the teaching and training of R extremely important.

With that being said, R's mastering has not been on student's to-do list until recent years. At the University of Toronto, up until fall in 2018, R's learning is not compulsory until third-year courses. Many students expressed surprise when they first see R's use in the classroom and are confused. The anxiety issue persists three years after the department made STA130 compulsory, which was an introduction to statistics and R. Studies show that the anxiety level affects students' performance in the classroom and has the potential for further investigation(Saade et al. 2017).

I am interested in the distribution of R anxiety levels among students. In this report, I will show the method to estimate the anxiety distribution among students using a sample I collected from the University of Toronto. I run a multilevel regression on my sample and then post-stratify the results using a simulated student census to get population estimates.

The results of the analysis can be useful in many ways. The university can periodically conduct this analysis to keep track of teaching results; generally, this method should be solid for similar reports in any other university. Students can use the information as a threshold to understand where they stand among their peers. It is also possible to study the effect of a treatment such as data camp using this approach, which potentially saves cost for the department.

In paper we run our analysis in `R` (R Core Team 2020). We also use the `tidyverse` package which was written by (Wickham et al. 2019).

*Code and data are available at: https://github.com/dding33/STA304-PS5.

Data

Multilevel regression and post-stratification require two data sets. We train a multilevel regression model using our sample data set(which is the smaller one), then apply the results to the second data set(usually a large data set like census) to mimic our population's behavior. In the context of this report, we want to estimate the anxiety score for all third-year statistics students at the University of Toronto.

The idea came from an example in mathexample studying anxiety level towards mathematics.

Sample from Survey

We use piazza and Quercus to distribute the survey organized on Google Forms[1] and to record the results. Naturally, the sampling frame comprises all third-year stats students who have access to the internet. The survey contains the following compulsory questions: -one qualification question to reduce sampling errors; -four demographic questions to post-stratify; -one question about anxiety level. We ask respondents to self-evaluate their anxiety level on a scale of 10 when asked to complete a task in R independently, where '1' represents not very anxious, and '10' represents feeling very anxious. We will refer to this by 'anxiety score' or 'anxiety level' for the rest of this report. In the end, we have an optional question where the respondent can express their opinion on how to reduce their anxiety towards R. For privacy reasons, the responses to this question are masked. Instead, there will be a summary of the responses later in the discussion section.

We are restricting the year of study because we want to see how the anxiety levels vary within a group of students with similar exposure to R; students in the same year have similar experiences both timewise and course-wise. It is intuitive that the more experience a student has(the closer he/she is towards graduation), the more familiar he/she becomes with R and thus has a lower anxiety score.

The total sample size is 48, from which 7 respondents answered 'no' to the qualification questions. That makes the sample size 41. Figure 1 shows the distribution of the programs from the sample. The coloring at the end of the bars indicates the respondent does not have any coding experience. We notice that this is a small part of the sample.

Figure 3 displays the anxiety score distribution from the raw data set. We see two prominent peaks in the distribution: around 3, which indicates the respondents do not feel very anxious, and around 7.5 indicates the respondents feel quite anxious. Most responses fall between these two peaks, with few respondents(3 out of 41) reports extreme anxiety scores towards 1 or 10. Looking at Figure 2, we notice that our sample does not contain any students with a cumulative GPA lower than C. This skewness means that our sample is biased; specifically, students with higher cgpa have a stronger incentive to answer the survey. We will adjust for this in the model by incorporating random effects.

We must point out that studies show the response biases for sensitive topics center are near zero, but the responses are unreliable or noisy(Marquis, Marquis, and Polich 1986). Since the cgpa and the anxiety score reveal information about students' academic behavior, we consider them sensitive topics. We follow steps from (Gelman 2019) using the `brms` package(Bürkner 2017, @brmtwo) to adjust for this. Further explanation will continue in the Model section.

Simulated Student Census

We simulate a census data set for all third-year stats students and use this as our post-stratification data. From admission information in 2017(the year that most third-year students in 2020 were admitted), we estimate our census's size to be 850. It contains five variables that describe each individual:

- Student_id;
- sex(2 levels);

[1]The link to the full survey: https://forms.gle/x4mxCLw6Hh8ecqmT7

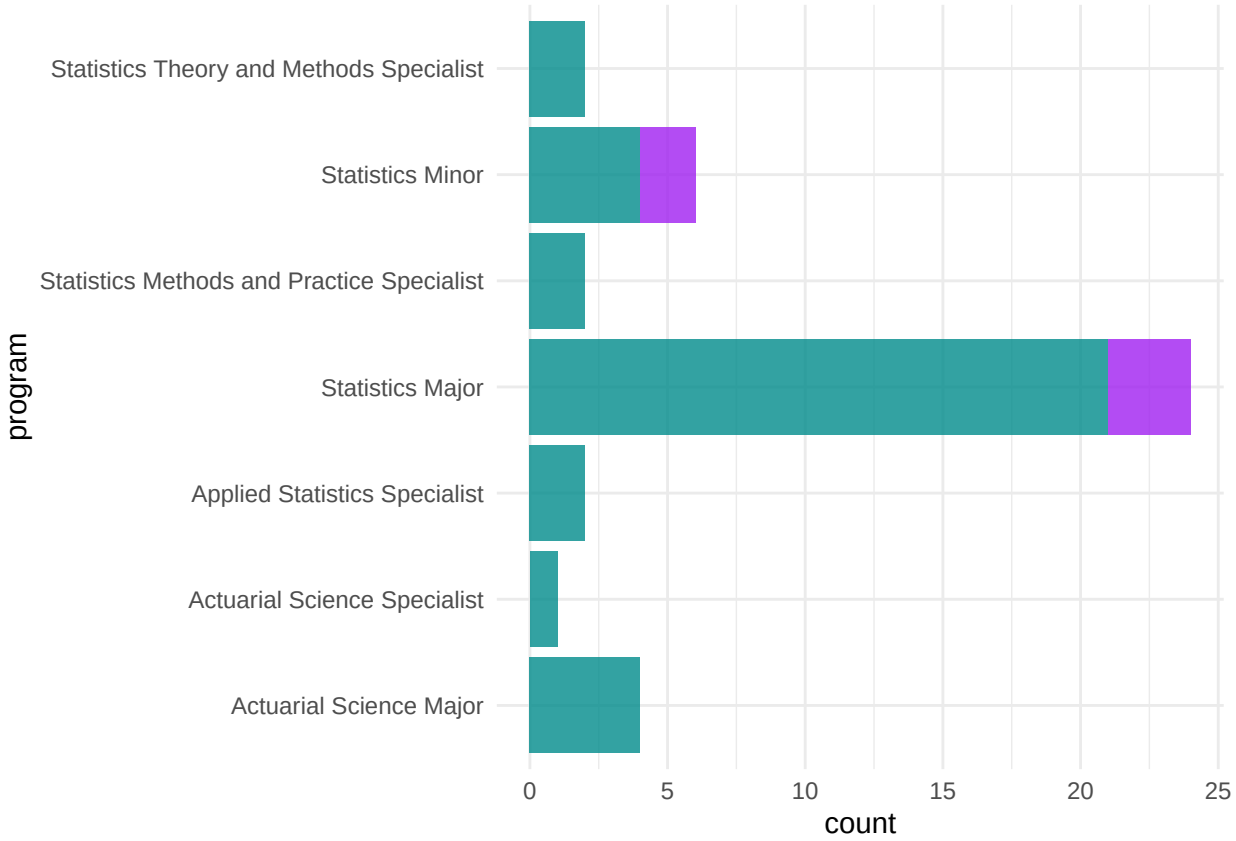

Figure 1: Distribution of Programs

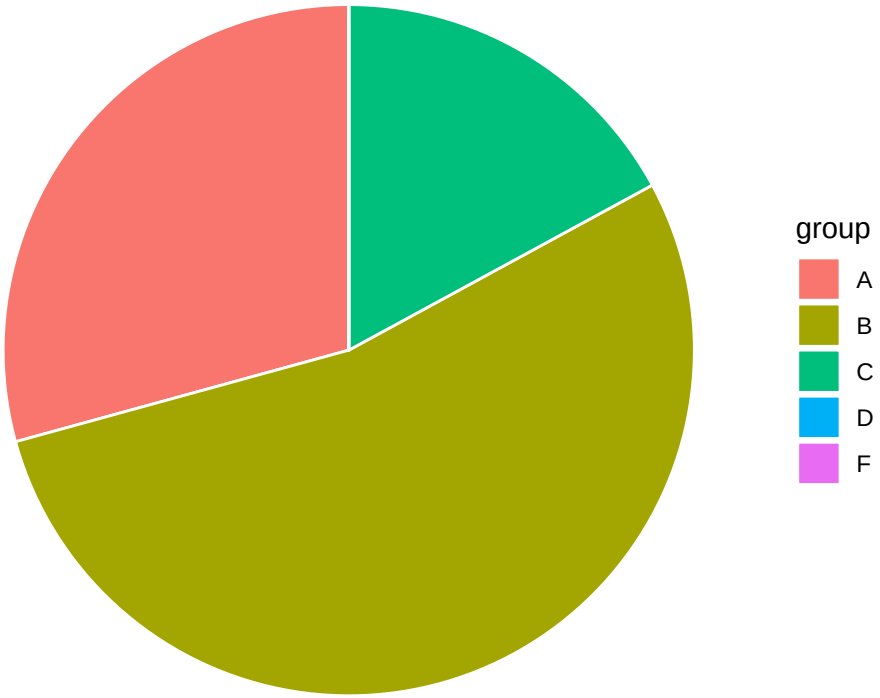

Figure 2: cgpa Distribution among Sample

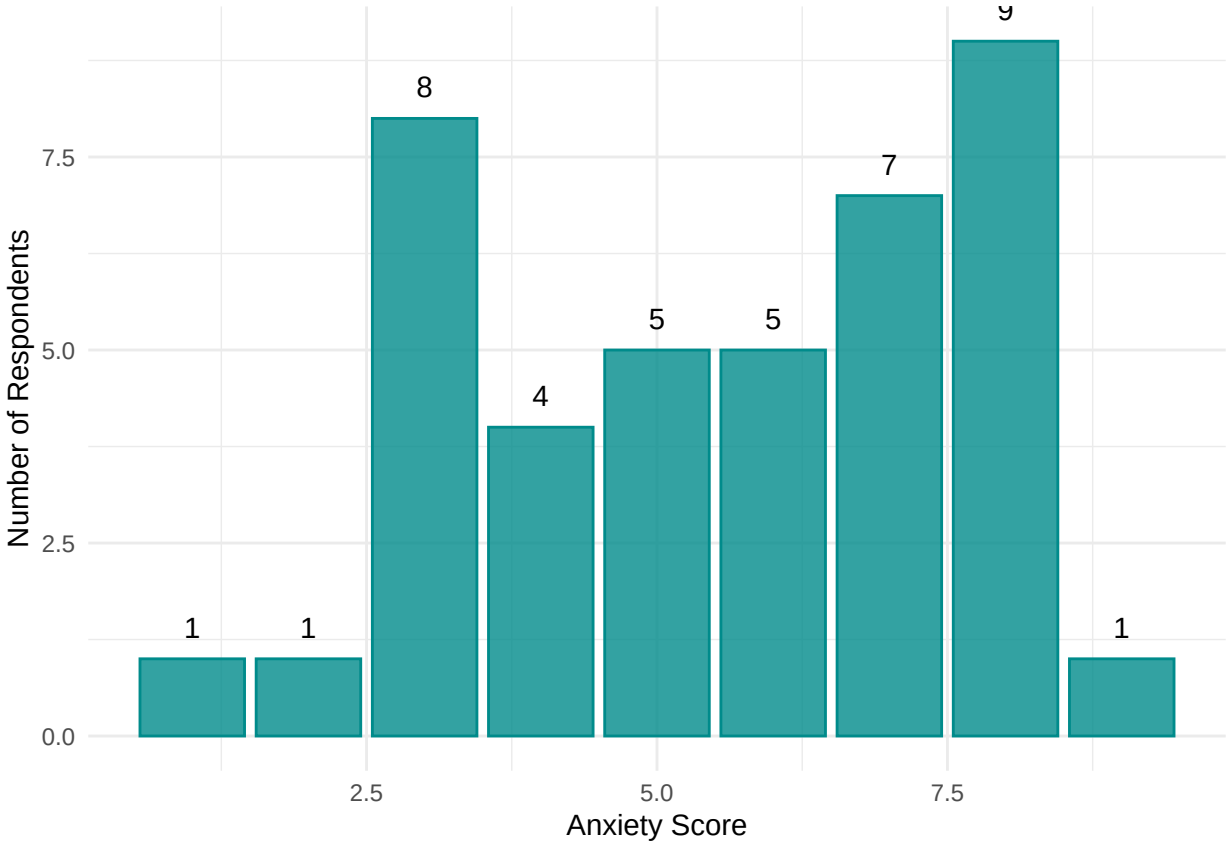

Figure 3: Distribution of R Anxiety Scores

- program(8 levels);
- cgpa(5 levels);
- conding_exp(2 levels).

We use ladder four to create post-stratification cells. They contain information about an individual to divide them into groups and identify each individual using these four variables. The four variables make $2*8*5*2 = 160$ possible cells. The distribution of each variable is simulated based on a rough estimation of the population. See fig 4 and 5 for a summary of the census. This report's results are not estimates of the University of Toronto's actual R anxiety distribution, even though the sample is collected from real respondents.

Based on the census, we developed a few prop data frames for post-stratification. We counted the number of individuals in each cell and saved it as cell_counts.

Figure 4: Program Distribution in Census

Model

MRP

We use MRP to predict the anxiety level distribution among our population. This method adjusts our estimation results by first fitting a multilevel regression model using the sample, then applying it to the post-strat data set to predict the population. Specifically, each individual is defined by his/her sex, program, cgpa, and whether he/she has previous experience with coding. For each individual in the census, we predict

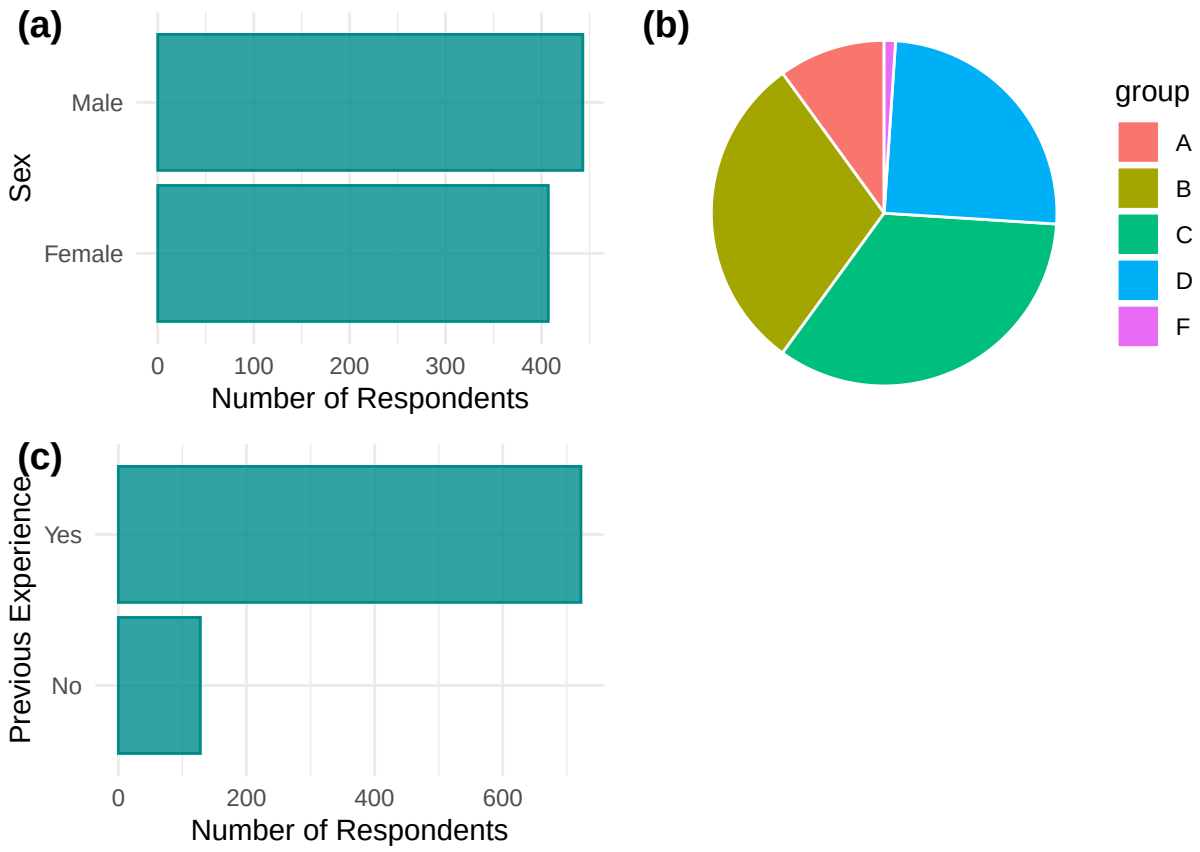

Figure 5: Distribtion of Other Variables in Census

that person's anxiety score using the previous model using `add_predict_draws()` from `tidybayes.` Then we aggregate the cell-level estimates up to the population level. Using y to represent the anxiety score,

$$\hat{y}_S^{PS} = \frac{\sum N_j \hat{y}_j}{\sum N_j} \tag{1}$$

We get our post-stratification estimates by equation (1). You can see that the key to an accurate estimate relies not only on how well the model fits the data but also on the level to which the census represents the population.

We get our post-stratification estimates by equation (1). You can see that the key to an accurate estimate relies not only on how well the model fits the data but also on the level to which the census represents the population. There is often a trade-off between the cells' division and the prediction results' stability [forcast]. In our case, the 160 possible cells divide the population very finely(5.3 persons in each cell on average), which is another reason for us to use MRP. Equation (2) shows the formula we use for the model, where β_{pro} represent the coefficient for the program beta, and d_{sex} represents the indicating variable for sex.

$$\hat{y} = \beta_0 + \beta_{sex} d_{sex} + \beta_{cgpa} x_{cgpa} + \beta_{pro} x_{pro} + \beta_{code} x_{code} + e \tag{2}$$

Model Validation

We perform k-fold cross-validation on the model fitted. This means refitting the model K times, leaving out one-kth of the original data each time. We are doing a 3-fold validation because our sample size is relatively small(41), and dividing it more than three times will lead to volatile results. We suggest you increase k as the sample size increases. The cross-validation estimates an average prediction error of 0.003, which indicates the model performance is not problematic.

Results

We use `add_predicted_draws()` to come up with estimations and their 95% confidence intervals. Figures 6-9 show the prediction results by different sex, program, cgpa, and coding experience, with the raw data results. We see that the MRP estimates produce a higher mean anxiety score comparing to the raw results. Specifically,

- MRP estimates substantially different mean anxiety scores for males and females, while there was no sign of this pattern among our sample. The estimated average for females is 6 and is 5 for the male. There is no significant difference between the interval of the groups;
- There is no significant difference in estimated mean anxiety scores between different programs. The confidence interval for the Actuarial Scitiest is longer than those of other programs, which means a broader range of anxiety levels within the program;
- Without any sample, the MRP predicts 5 to be the mean anxiety score for students with a cumulative GPA F. This group also has the widest confidence interval. Students with a cumulative GPA B seems more anxious towards the use of R than any other grades group, with the highest mean anxiety score and the narrowest confidence interval; its upper boundary is very close to 10, the highest possible anxiety score;
- Students with some previous coding experience are estimated to have a lower mean anxiety score than those who are new to programming. The experienced group also has a much smaller confidence interval, which indicates more stability. The inexperienced group's upper boundary almost reaches 10, but its lower boundary is close to the lower boundary for the experienced group.

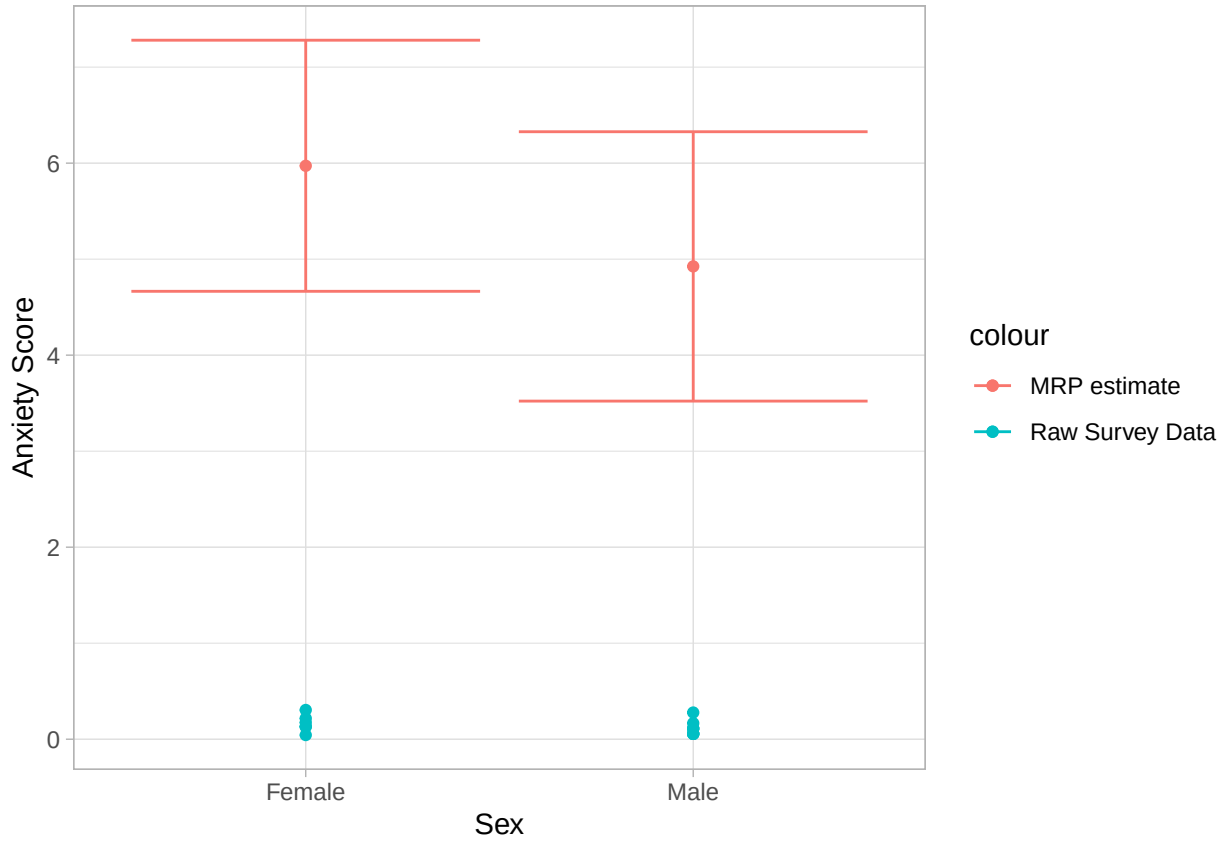

Figure 6: MRP estimates vs Raw data in different sex groups

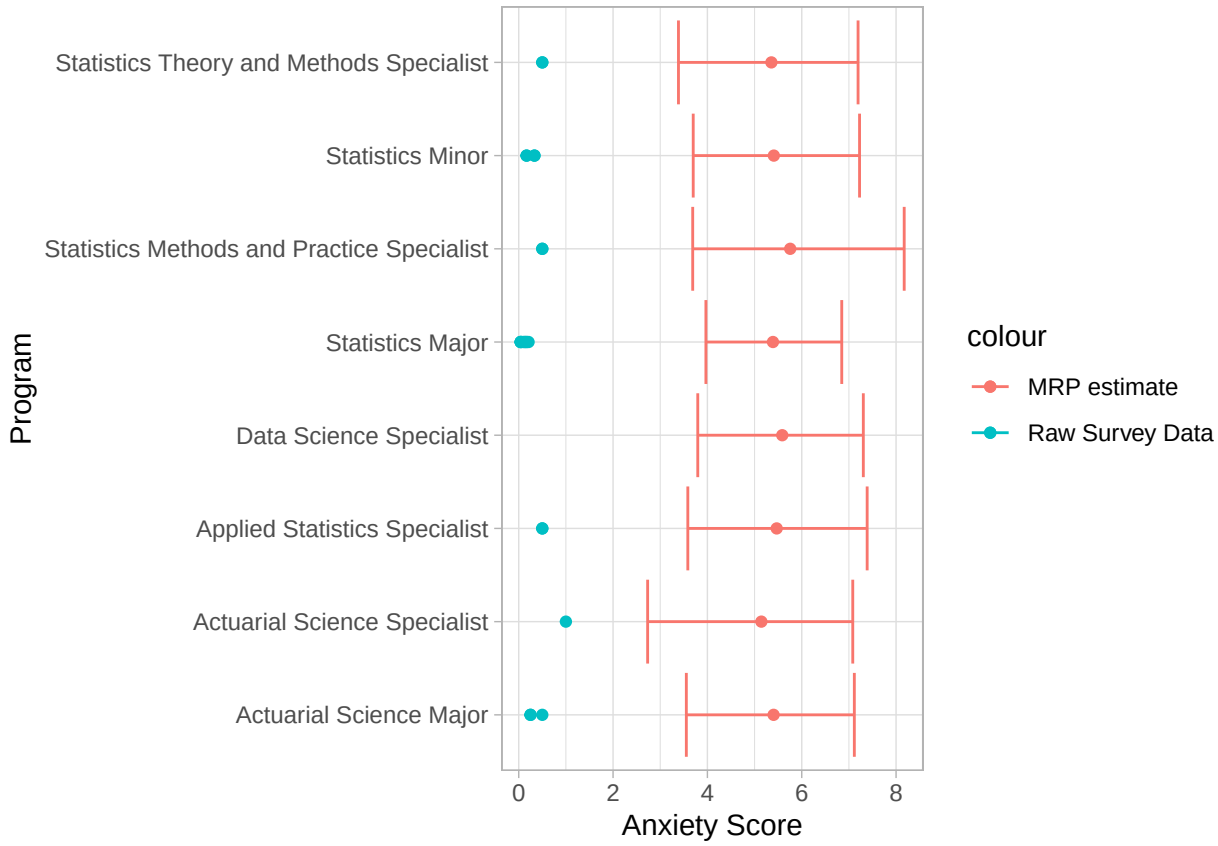

Figure 7: MRP estimates vs Raw data in different programs

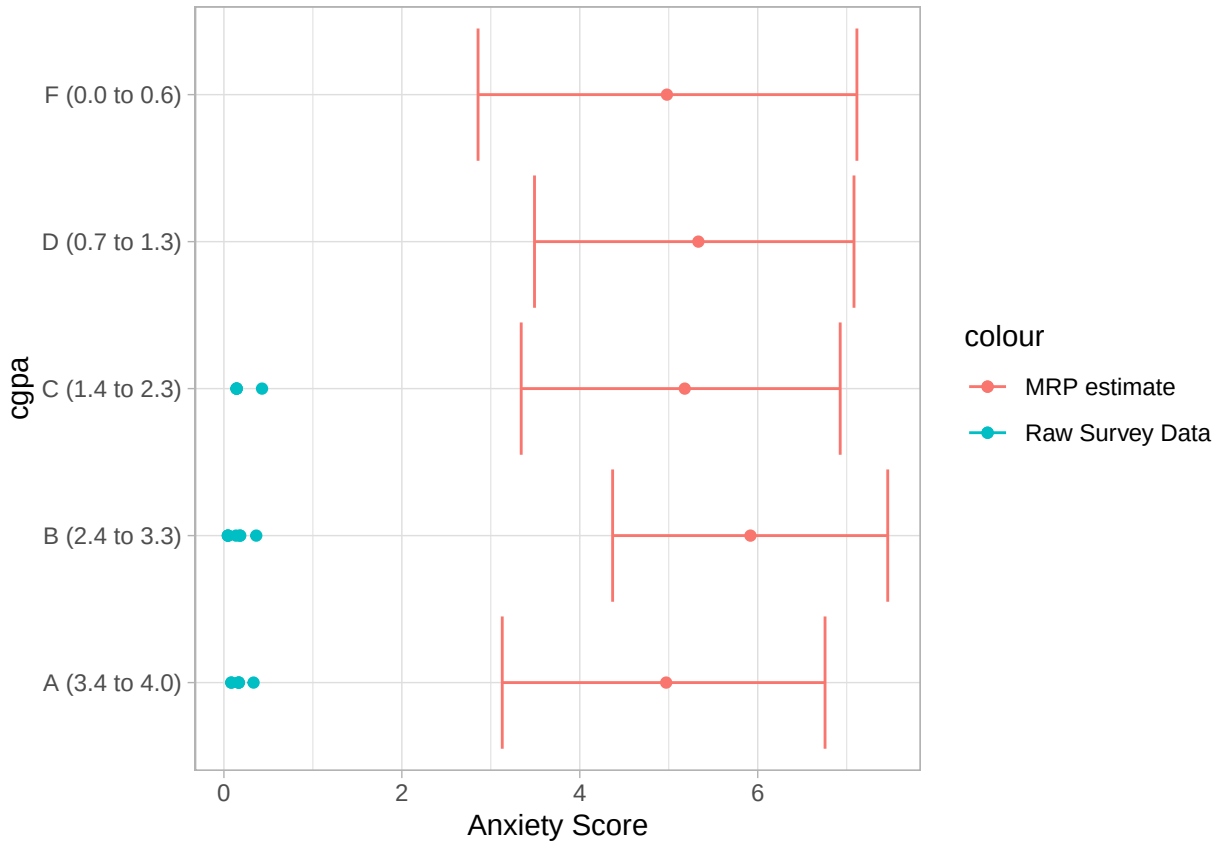

Figure 8: MRP estimates vs Raw data within different cgpa groups

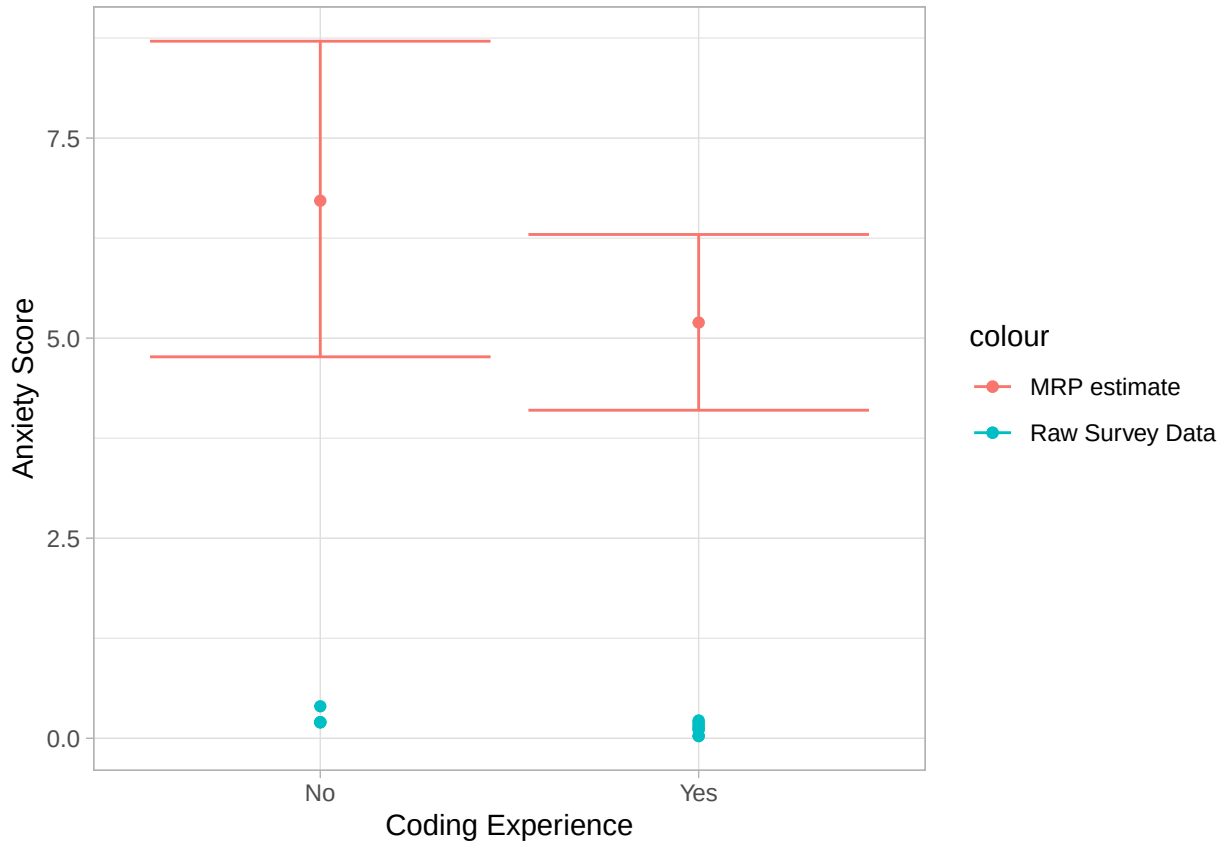

Figure 9: MRP estimates vs Raw data with different coding experiences

Discussion

Limitation and Future Researches

(Hanretty 2019) states that MRP works well adjusting for biased samples only if the under/over-represented variables are present in the post-stratification data set. With that being said, to get a more reliable prediction result, future analysis can contain a pre-analysis which has more variables and use stepwise regression to select variables that contribute to the accuracy of the model.

Appendix

References

Bürkner, Paul-Christian. 2017. "brms: An R Package for Bayesian Multilevel Models Using Stan." *Journal of Statistical Software* 80 (1): 1–28. https://doi.org/10.18637/jss.v080.i01.

———. 2018. "Advanced Bayesian Multilevel Modeling with the R Package brms." *The R Journal* 10 (1): 395–411. https://doi.org/10.32614/RJ-2018-017.

Gelman, Andrew. 2019. "Know Your Population and Know Your Model: Using Model-Based Regression and Poststratification to Generalize Findings Beyond the Observed Sample."

Hanretty, Chris. 2019. "An Introduction to Multilevel Regression and Post-Stratification for Estimating Constituency Opinion." *Political Studies Review* 18 (July): 147892991986477. https://doi.org/10.1177/1478929919864773.

Marquis, Kent H., M. Susan Marquis, and J. Michael Polich. 1986. "Response Bias and Reliability in Sensitive Topic Surveys." *Journal of the American Statistical Association* 81 (394). [American Statistical Association, Taylor & Francis, Ltd.]: 381–89. http://www.jstor.org/stable/2289227.

Nelder, John A. 1999. "From Statistics to Statistical Science." *Journal of the Royal Statistical Society. Series D (the Statistician)* 48 (2). [Royal Statistical Society, Wiley]: 257–69. http://www.jstor.org/stable/2681191.

R Core Team. 2020. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Saade, Raafat, Dennis Kira, Tak Mak, and Fassil Nebebe. 2017. "Anxiety & Performance in Online Learning." In, 147–57. https://doi.org/10.28945/3736.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain Francois, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.