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

Yiqu Ding

2020-12-22

Abstract

In this report, I looked 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 differ from the raw data's average anxiety scores.

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 2018, R's learning is not compulsory for first-year stats students. 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).

This report focuses on the distribution of R anxiety levels among students. We 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). See the complete list of packages used in the appendix.

*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 Gelman (2019) 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:

- 1 qualification question to reduce sampling errors;
- 4 demographic questions to post-stratify;
- 1 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 restrict 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 of study 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: at around 3, which indicates the respondents do not feel very anxious, and at 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)(Bürkner 2018) to adjust for this bias.

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

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

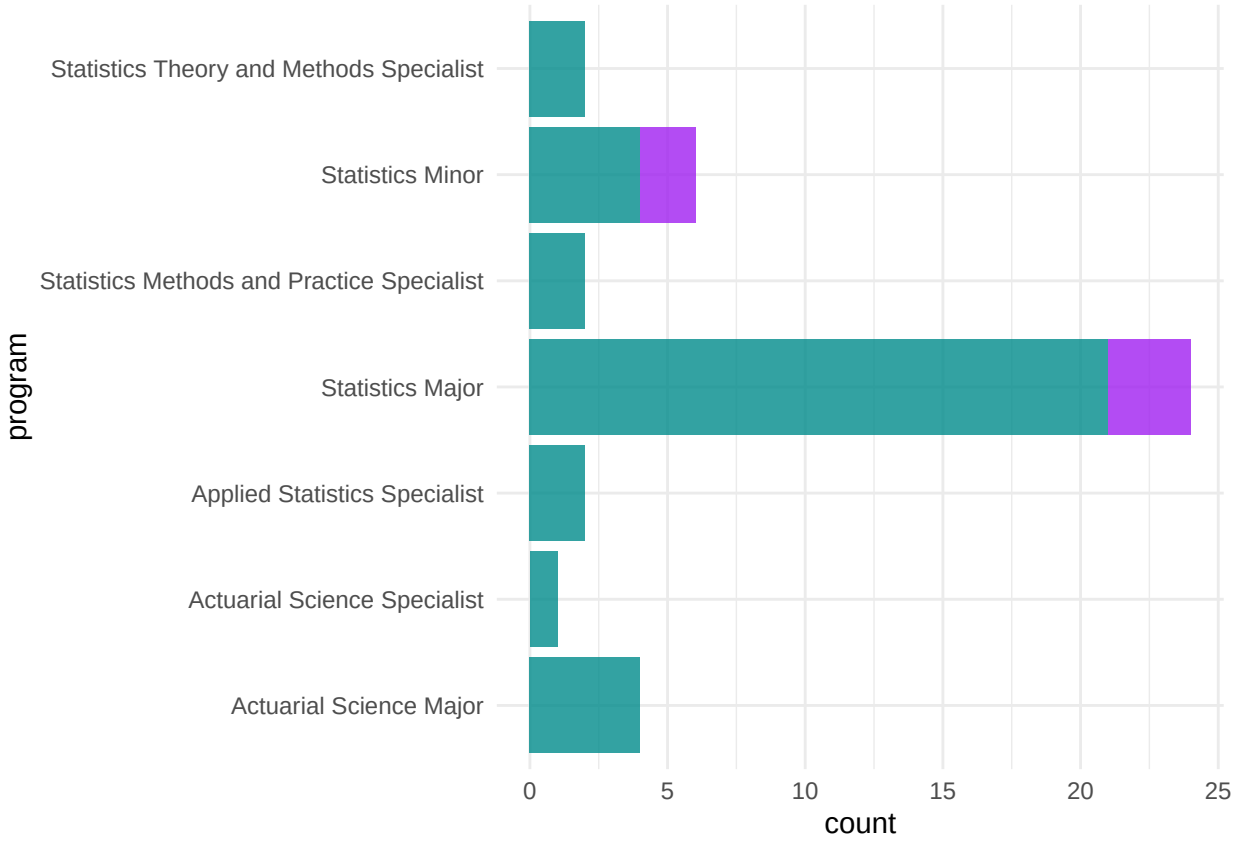

Figure 1: Distribution of Programs

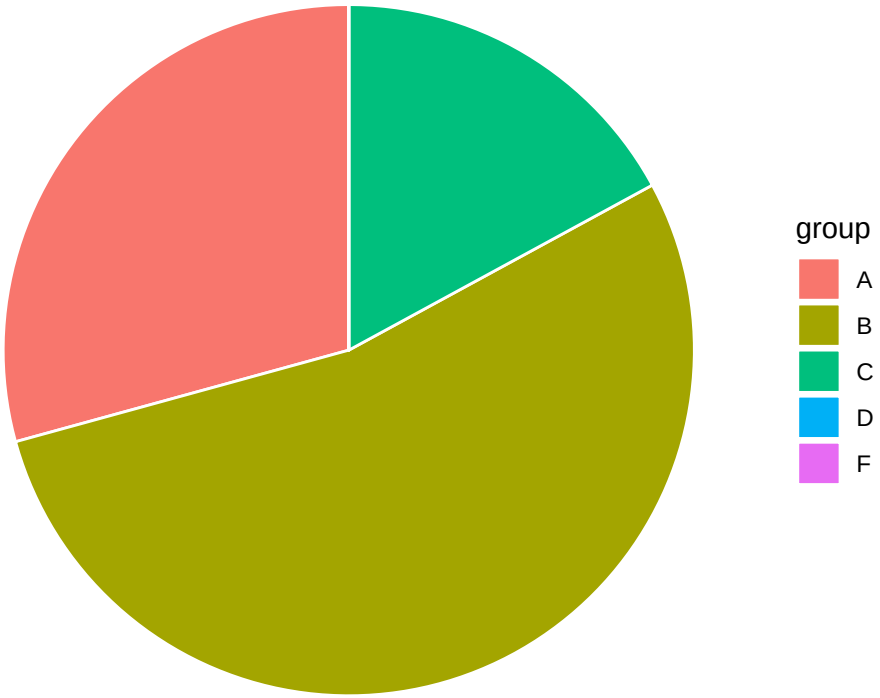

Figure 2: cgpa Distribution among Sample

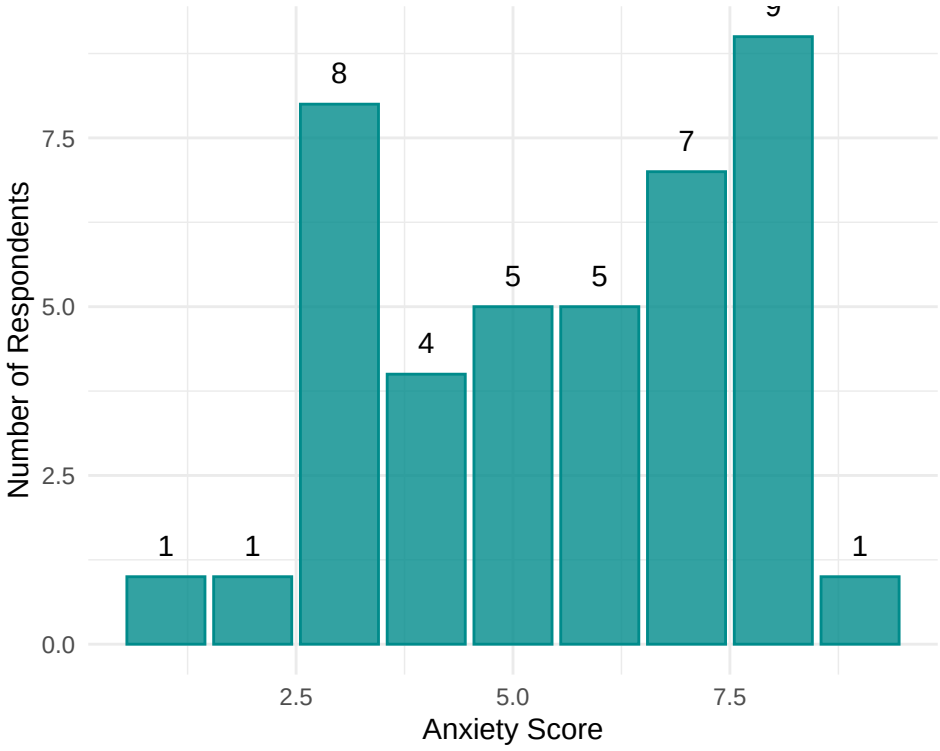

Figure 3: Distribution of R Anxiety Scores

estimate our census's size to be 850. It contains five variables that describe each observation, that is, each individual:

- student_id
- sex(2 levels)
- 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 them using these four variables. The four variables make $2 * 8 * 5 * 2 = 160$ possible cells. Each variable's distribution is simulated using random sampling with proportions for some variables and stabilized using a seed. We base the proportions on a rough estimation of the population. The proportions of females and males are randomly selected; the resulted ratio seems reasonable(around 1:1). We fail to find relative information on the distribution of programs within a department; therefore, each program's proportion is based on a completely arbitrary list of proportions. We estimate the most students are getting a cgpa of B or C, considering the year of study and some course averages available. Given the existence of STA130, a compulsory course for first-year stats students since 2017, we expect most of the population to have some sort of experience with codes. However, students may transfer into the program after the first-year hence have not taken this course. Some students pursuing a minor in statistics might not have any experience with codes as well. Therefore, we assume 85% of the population has some experience with codes and 15% are inexperienced. See fig 4 and 5 for a summary of the census.

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. We then created data frames for each variable to record the proportions of each type of individual in the population.

Again, using a simulated census means this report's results are not estimates of students in the University

of Toronto's actual R anxiety distribution, even though the sample is collected from real respondents. This report aims to show how the method of MRP can be applied in this particular context and how we would have interpreted it if we use the actual census.

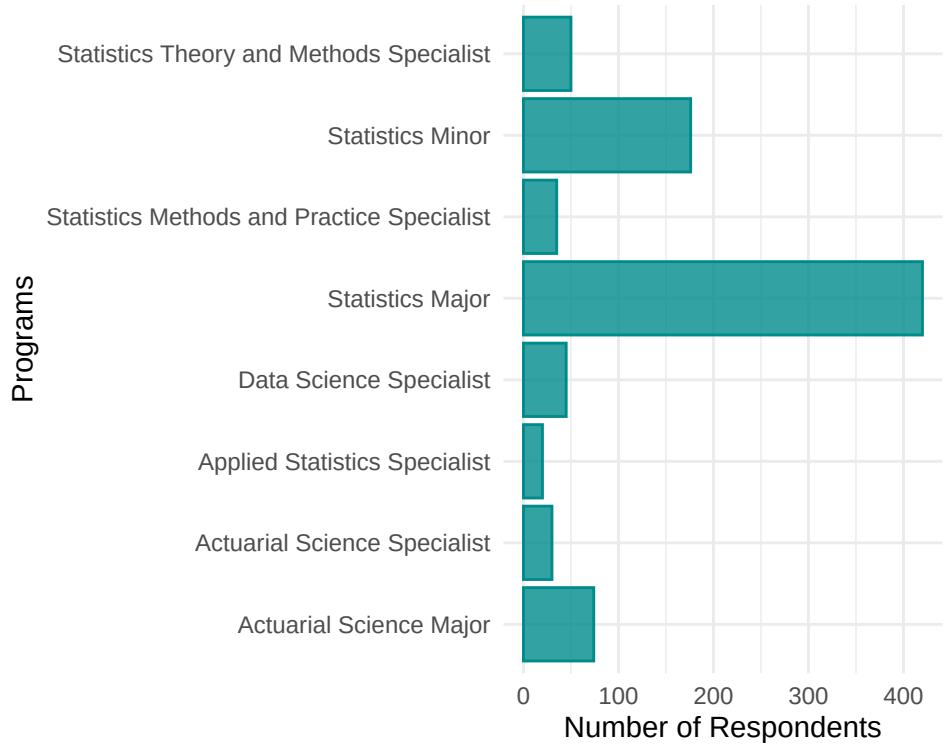

Figure 4: Program Distribution in Census

Model

MRP

We use multilevel regression with post-stratification 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 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, we can estimate the anxiety score for any subpopulation:

$$\hat{y}_{sub}^{PS} = \frac{\sum N_j \hat{y}_{j\, j \in J_{sub}}}{\sum N_{j\, j \in J_{sub}}} \tag{1}$$

We get our post-stratification estimates by equation (1), where J_s are all the cells that are in the subpopulation and $\hat{y}_{sub}^{PS}$ is based on our multilevel regression model. 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 (Wang et al. 2015). In our case, the 160 possible cells divide the population finely(5.3 persons in each cell on average), which

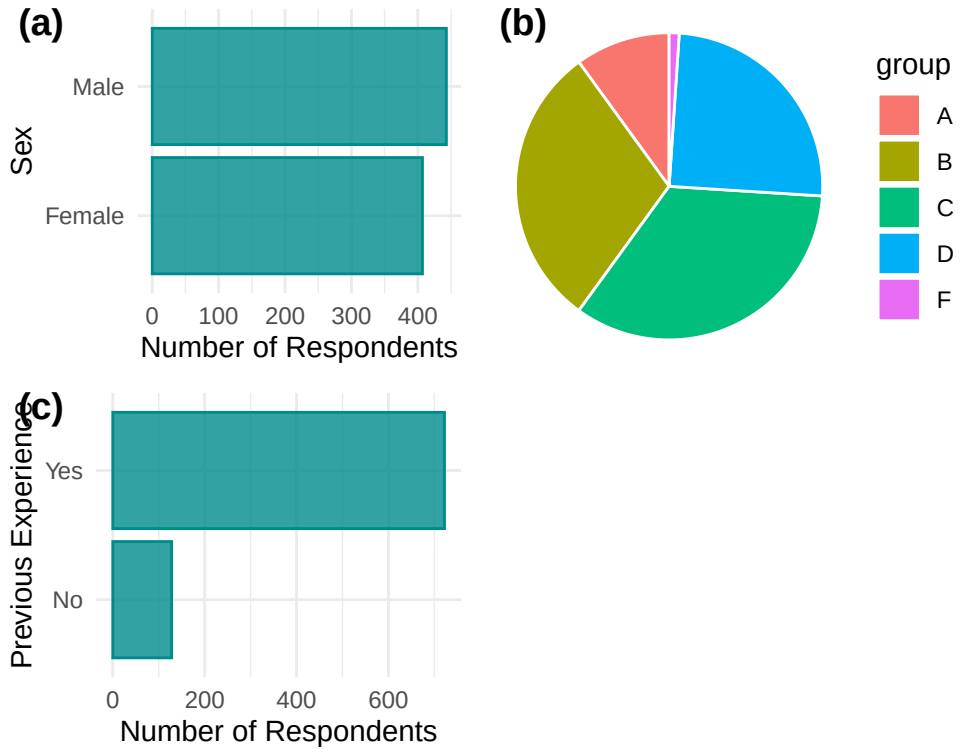

Figure 5: Distribtion of Sex, Cgpa and Coding Experience in Census

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 fitted model. 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 to 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 different mean anxiety score for each subpopulation group comparing to the raw results. Specifically:

- MRP estimates substantially different mean anxiety scores for males and females. The estimated average for females is 6 and is 5 for the male. There is no significant difference between the groups' interval, the MRP estimator for males is lower than raw data, and for females, it is higher than raw

data. This means the gap between males and females in the population is not as extreme as it is in our sample;

- There is no significant difference in estimated mean anxiety scores between different programs. The confidence intervals for the actuarial specialist and the statistics methods and practise specialist are longer than those of other programs, which means a broader range of anxiety levels within the program. The MRP estimate for the three programs is higher than the raw data average, and the MRP estimate for the rest five programs are lower than the raw data average. The Statistics Methods and Practise specialist also has the highest upper bound for its 95% Confidence interval of the estimate;
- Without any sample, the MRP predicts 5 to be the mean anxiety score for students with a cumulative GPA F and 5.4 to be the mean anxiety score for students with a cumulative GPA D. The F group also has the widest confidence interval. Students with a cumulative GPA of B seems more anxious towards the use of R than any other grade 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. The MRP estimate is lower than the raw data average for group B and higher than the raw data for group A and C;
- Students with some previous coding experience are estimated to have a significantly 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. Both MRP estimates are lower than the raw data average, and the difference became minimal for the experienced group.

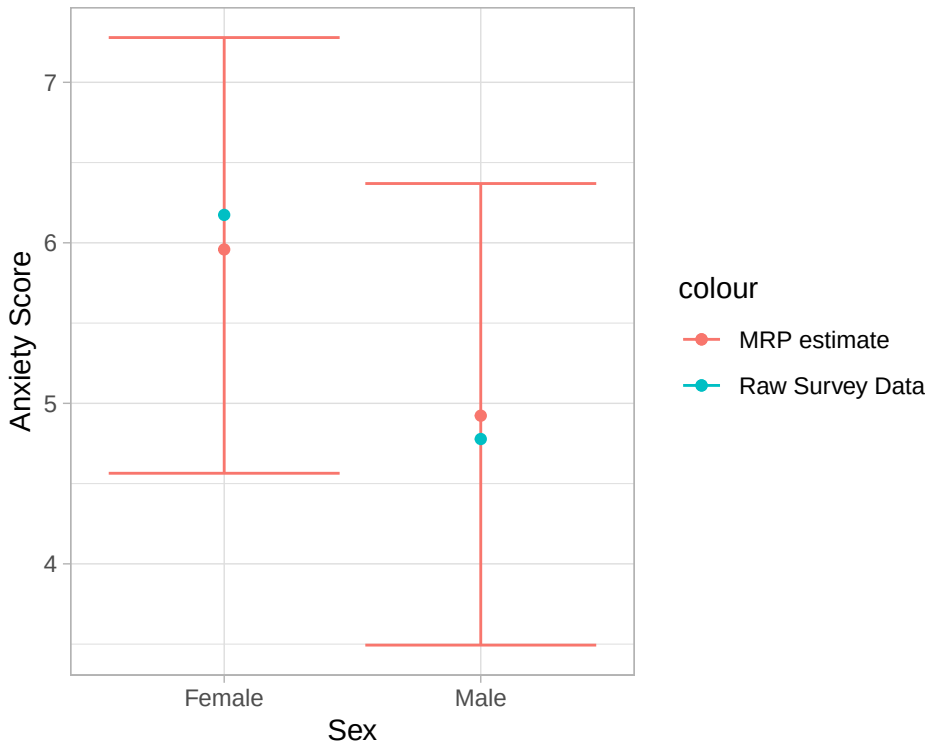

Figure 6: MRP estimates vs Raw data in different sex groups

Discussion

We observe that the MRP estimates tend to stabilize the average anxiety score; this is especially obvious in Figure 6 and Figure 8, where it "pulls" the more extreme scores towards the middle. Its estimations have

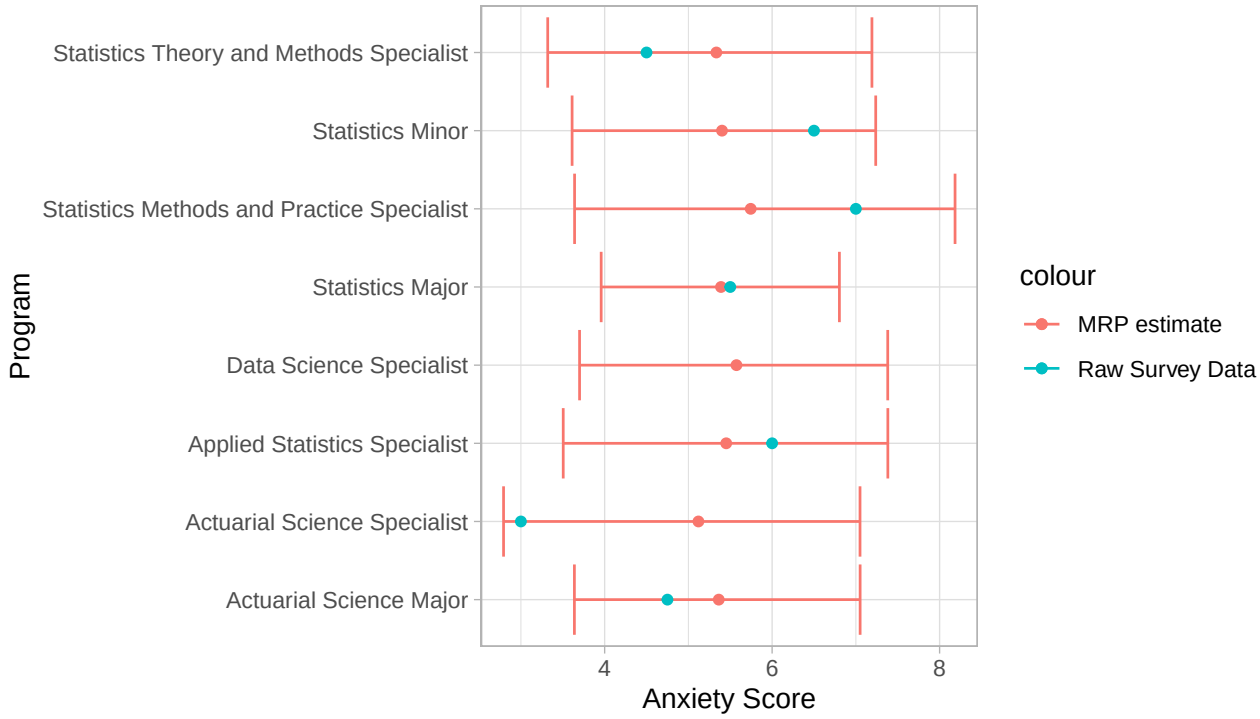

Figure 7: MRP estimates vs Raw data in different programs

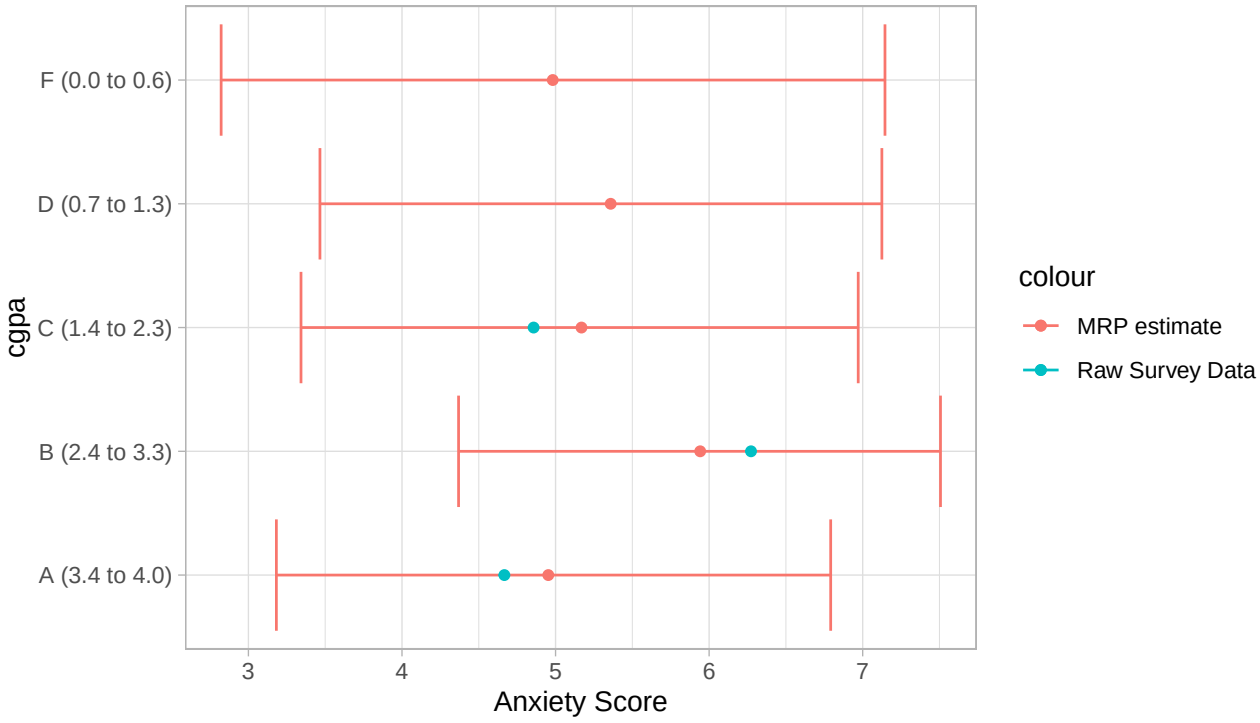

Figure 8: MRP estimates vs Raw data within different cgpa groups

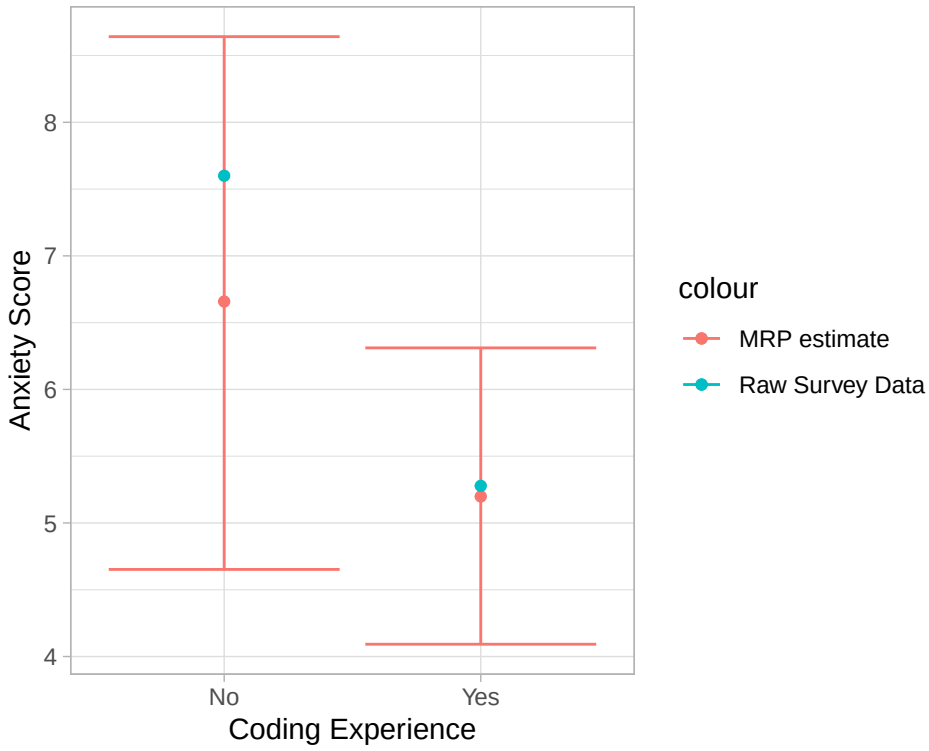

Figure 9: MRP estimates vs Raw data with different coding experiences

no significant difference for different programs, even though the raw data averages vary. This is likely due to the small sample size where extreme answers have a strong inference on the average; once we apply the model to the population, the difference tends to cancel out.

For different grade groups, group B particularly stands out. This pattern indicates that the group of students getting B might be facing more pressure towards R than students earning other GPA in general. The causality between their exceptionally high anxiety level and the GPA is worth exploring. One explanation for this is: students with an A uses R very well and do not feel too stressed; at the same time, the mastering of R is not the primary concern for students with C, D or F. While students with a B still care about this skill, they are more stressful than A students because they are relatively less familiar with R.

Even "pulled", the MRP estimates different anxiety levels on average for males and females. This difference is not uncommon in many so-called male-dominant industries like software development and IT. For example, studies show that female developers are hesitant to explore the opportunities to contribute to new projects even when they possess the competence to make valuable contributions(Wang, Wang, and Redmiles 2018). Female students may feel less confidence with their R skills for similar reasons.

In conclusion, the anxiety level within programs does not differ significantly. But for sex, it does. We suggest using sex-specific treatments to resolve this. We suggest paying attention to mental health and focusing on building confidence among students to reduce the higher than other anxiety levels for students earning B. A study regarding the relationship between the anxiety level and GPA states depression is a significant predictor of lower GPA and a higher probability of dropping out, particularly among students who also have a positive screen for an anxiety disorder(Eisenberg, Golberstein, and Hunt 2009); even though anxiety towards R does not necessarily imply anxiety disorder, it could still afftect studnts' academic behaviour and have a negative effect on maintaining students' mental health. Students with coding experience are significantly less anxious than those who have not been using R, which indicates some introductory course could help reducing students' anxiety level towards using R. Adding a specific introductory course for R like CSC108 is for python also appeared as answers to student's solution towards R anxiety. However, we expect the inexperienced group's size to drop in the next few years in the University of Toronto. As older students

graduate, the new generation of statistics students should all have taken STA130, which means they will have some experience towards R before they are asked to complete more complicated tasks with R.

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, therefore to comfirm the use of MRP.

Another issue with this report is that the prediction relies on a simulated census, which is in no way a true representation of the students actually studying at the University of Toronto. Even though the multilevel regression model is based on the actual response, we cannot mimic the population with this simulated census. The sample is relatively small, which means when we divide them into post-stratification cells we face the trade-off between the number of cells and the number of individuals within each cell. In current stage of the study, each small cells is referencing cells that are very similar to it, and this can be eliminated by increasing the sample size. For the results to be meaningful, future researchers should consider increasing the size of the sample, and accordingly change K in the cross-validation section to examine the model.

Treatment Analysis

An important application of this method is to estimate the effect of treatments on reducing the anxiety score. Facing high anxiety estimation, the university naturally would want to reduce this anxiety among students. Implementation of such actions is due to be costly because of the number of students who need to be involved and the uncertainty of the effects. Using MRP, we can model the difference between pre and post-intervention groups without necessarily carrying out the implementation of the population(Gelman 2019). We would only need a small sample of students who accepts the intervention and model the results based on the sample. We then compare the estimation with the pre-intervention group to see whether the treatment is effective or worth the spending. The process is now shown in this report because we do not want people to assume what their anxiety score will be after the treatment; it is quite tricky and often not accurate to do so.

Student's Opinion

We got 29 responses from the optional question on the survey: "What would ease your worries about R?" and the responses can be summarized into three categories:

- Immediate Help. Students claim that they find it helpful to get immediate help with error codes, either from the teaching team or from fellow students. Some students mentioned more TA office hours.
- More examples/instructions. Students often find it hard to interpret a project given in words. They point out that if more examples can be given, they know more about what is expected and should be less anxious.
- A better introduction to R. Students often find a gap between what the teaching team expects from first-time R users and their understanding of R. Some suggested opening up an introductory course to R similar to CSC108.

Appendix

Platforms

The survey and its responses were organized using Google Forms[2].

The survey was distributed on Quercus and Piazza[3] thanks to professor Alexandar and professor Caetano at the Univeristy of Toronto.

This reported was created using(Allaire et al. 2020; Xie, Allaire, and Grolemund 2018)

[2]https://www.google.com/forms/about/
[3]https://piazza.com

References

Allaire, JJ, Yihui Xie, Jonathan McPherson, Javier Luraschi, Kevin Ushey, Aron Atkins, Hadley Wickham, Joe Cheng, Winston Chang, and Richard Iannone. 2020. *Rmarkdown: Dynamic Documents for R*. https://github.com/rstudio/rmarkdown.

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.

Eisenberg, Daniel, Ezra Golberstein, and Justin B Hunt. 2009. "Mental Health and Academic Success in College." *The BE Journal of Economic Analysis & Policy* 9 (1). De Gruyter.

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.

Kassambara, Alboukadel. 2020. *Ggpubr: 'Ggplot2' Based Publication Ready Plots*. https://CRAN.R-project.org/package=ggpubr.

Kay, Matthew. 2020. *tidybayes: Tidy Data and Geoms for Bayesian Models*. https://doi.org/10.5281/zenodo.1308151.

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.

Wang, Wei, David Rothschild, Sharad Goel, and Andrew Gelman. 2015. "Forecasting Elections with Non-Representative Polls." *International Journal of Forecasting* 31 (3): 980–91. https://doi.org/https://doi.org/10.1016/j.ijforecast.2014.06.001.

Wang, Zhendong, Yi Wang, and David Redmiles. 2018. "Competence-Confidence Gap: A Threat to Female Developers' Contribution on Github." In *Proceedings of the 40th International Conference on Software Engineering: Software Engineering in Society*, 81–90. ICSE-Seis '18. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3183428.3183437.

Wickham, Hadley. 2016. *Ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org.

———. 2020. *Forcats: Tools for Working with Categorical Variables (Factors)*. https://CRAN.R-project.org/package=forcats.

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.

Wickham, Hadley, Jim Hester, and Romain Francois. 2018. *Readr: Read Rectangular Text Data*. https://CRAN.R-project.org/package=readr.

Xie, Yihui, J.J. Allaire, and Garrett Grolemund. 2018. *R Markdown: The Definitive Guide*. Boca Raton, Florida: Chapman; Hall/CRC. https://bookdown.org/yihui/rmarkdown.