A Multivariate Logistic Regression of Gender for University of Waterloo Students

Allen Li

22 Dec 2020

Github

https://github.com/youhjjhhhjj/5howqCgU3f

Abstract

For the purpose of investigating any potential gender disparity in university institutions, the 2016/2017 University of Waterloo student headcount data was used to build a multivariable logistic regression model that relates an individual's gender to their year of study, faculty group, co-op/attendance/work-term status, and visa status. The resulting model shows that some factors are significantly correlated with gender, which suggests that men and women may not be on an even playing field in their university studies. An analysis is carried out that relates the results to established knowledge regarding gender differences in post-secondary institutions.

Keywords

gender, inequality, university, post-secondary, University of Waterloo

Introduction

The gender diversity of post-secondary students has been a hot topic for the past couple of years. It is well documented that women are underrepresented in various fields of study, especially STEM, and men are underrepresented in fields such as education, health, and welfare, along with being proportionally underrepresented in post-secondary as a whole[2]. Identifying and addressing these inequalities would be instrumental in removing barriers and increasing occupational diversity, which would result in a more equitable outcome for everyone.

In order to address this issue, it is necessary to first investigate the factors that are potentially correlated with a university student's gender. To this end, a multivariable logistic regression will be carried out with gender as the dependent variable and year of study, faculty group, co-op status, attendance, term type, and visa status as predictors. Note that this regression is meant to investigate correlation, it does not claim that any factor causes a student's gender.

The University of Waterloo student demographics data set will be used for the logistic regression. In the methodology section, the specifics of the data is discussed and the logistic regression model is described. In the results section, the details of the fitted logistic regression model are provided, along with additional analysis. The discussion section goes over the results in more detail and provides interpretations, and discusses weaknesses of the analysis along with future steps that could be taken.

Methodology

Data

The source of the data is the IAP Count Data Database for University of Waterloo students. The data originates from the student registration system, so it contains the information of every registered student at the University of Waterloo each term. To avoid double-counting, a single financial year was chosen for the analysis, the financial year of 2016/2017, because it is the most recent year for which gender is included as a variable. The data originally came in the form of a contingency table, so it was converted into a data frame with a row for each student. There were only 7 cases where the gender was neither male nor female, so those cases were removed to make the data more suitable for logistic regression. During data analysis, it was discovered that only co-op students could could have a work term, since regular students were restricted to academic terms. Similarly, only academic terms could be part-time, and thus there were 0 instances of part-time work terms. In light of this, the columns Coop.Regular, Attendance, and WorkTerm were combined into a new column, Coop.Attendance.WorkTerm, with 5 levels. This new variable is graphed in figure 2. Furthermore, it was discovered that there was significant correlation between the variables Career, Program.Level, and Study.Year, shown in tables 1 and 2, so it was decided that only Study.Year would be included in the model. Finally, a binary variable isFemale was added, which encodes for whether the student's gender is female. The analysis aims to provide insight into universities as a whole, especially those in Canada, so the target population is all university students in Canada. The both the frame and sample populations are all University of Waterloo students, since the data set includes the information of every registered student. This is a major strength of the data set, since it will be free from sampling biases. A weakness of the data set is that it is exclusive to Waterloo, so any conclusions made would be more valid for universities that are culturally similar to Waterloo, and less valid for those that are different.

Figure 1: Bar Plot of Gender, Faculty Group, and Visa Status

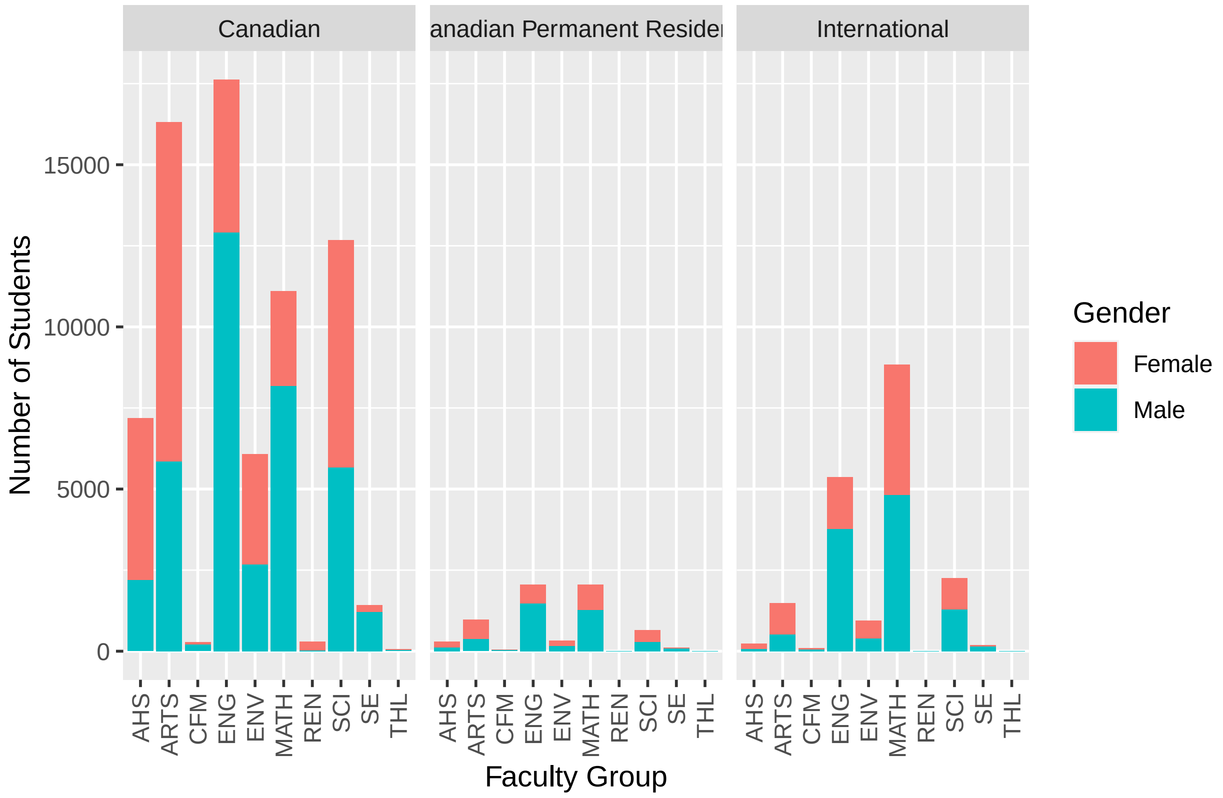
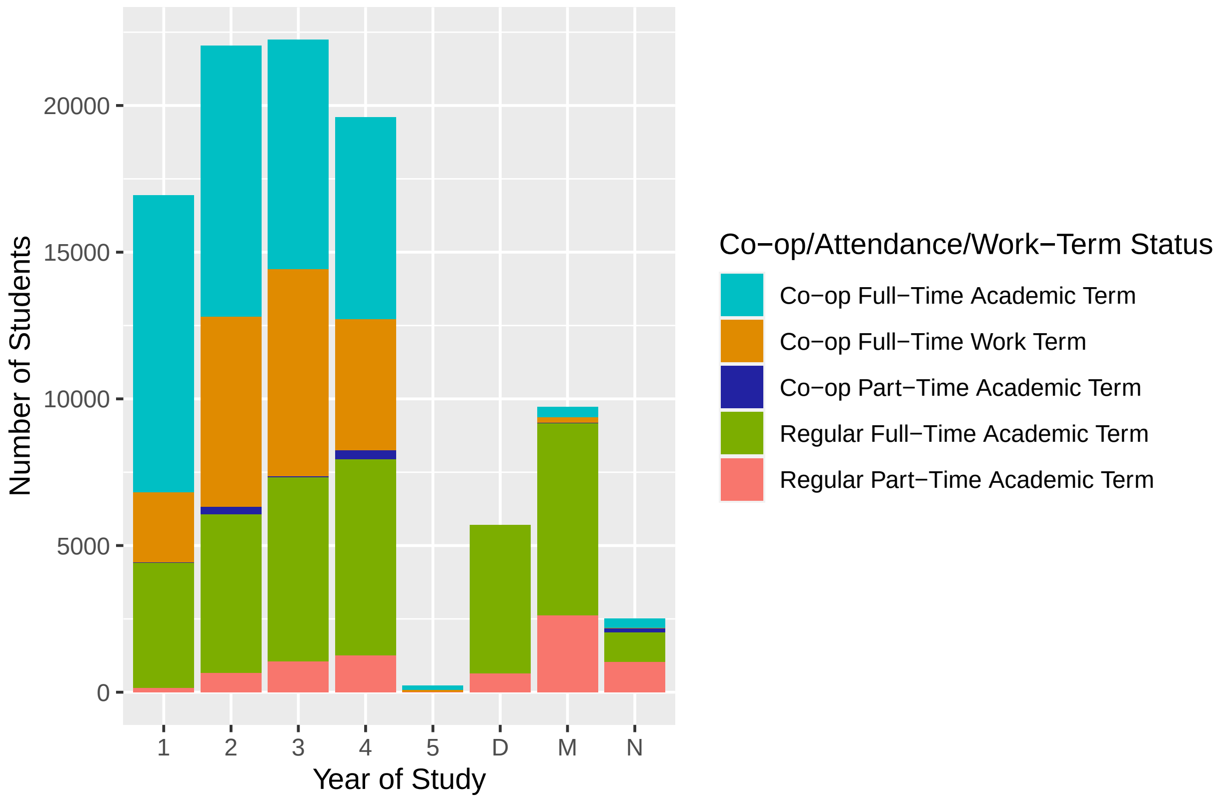


Figure 2: Bar Plot of Year of Study and Co−op/Attendance/Work−Term St



	1	2	3	4	5	N
Bachelors	16945	22047	22238	19599	237	0
Non-Degree	0	0	0	0	0	2440

Table 1: Table of Program Level and Study Year for Undergraduate Students

	D	M	N
Doctoral	5703	57	0
Masters	2	9665	0
Non-Degree	0	0	68
Qualifying	0	0	1

Table 2: Table of Program Level and Study Year for Graduate Students

Model

A multivariable logistic regression is used because this is most appropriate for a regression with multiple independent variables and a binary dependant variable. The independent variables are the student's year of study, faculty group, co-op/attendance/work-term status, and visa status, which were chosen because these factors are crucial considerations in one's education career, so it is important to determine whether or not they are correlated with gender. The original 4 categorical predictors have been decomposed into 26 separate binary variables (dummy variables) for the model. Note that the first dummy variable from each variable is excluded in order to reduce multicollinearity. The formula for the multivariable logistic model is given by

$$log(\frac{p}{1-p}) = \beta_0 + (\beta_{11}d_{11} + ... + \beta_{17}d_{17}) + (\beta_{21}d_{21} + ... + \beta_{29}d_{29}) + (\beta_{31}d_{31} + ... + \beta_{34}d_{34}) + (\beta_{41}d_{41} + \beta_{42}d_{42})$$

where p is the predicted probability of the student being female, β_0 is the log y intercept representing a student encompassing every reference level, which is a first-year Canadian citizen in the AHS faculty doing a regular full-time academic term, d_{1i} represents the i'th category of Study.Year, d_{2i} represents the i'th category of Faculty.Group, d_{3i} represents the i'th category of Coop.Attendance.WorkTerm, and d_{4i} represents the i'th category of Visa.Status. Each corresponding β represents the change in log odds for a deviation from the reference level in that category. The decision was made to not include interaction terms because there are no two variables that are expected to be related, thus avoiding needless complexity. The model is created using R.

Results

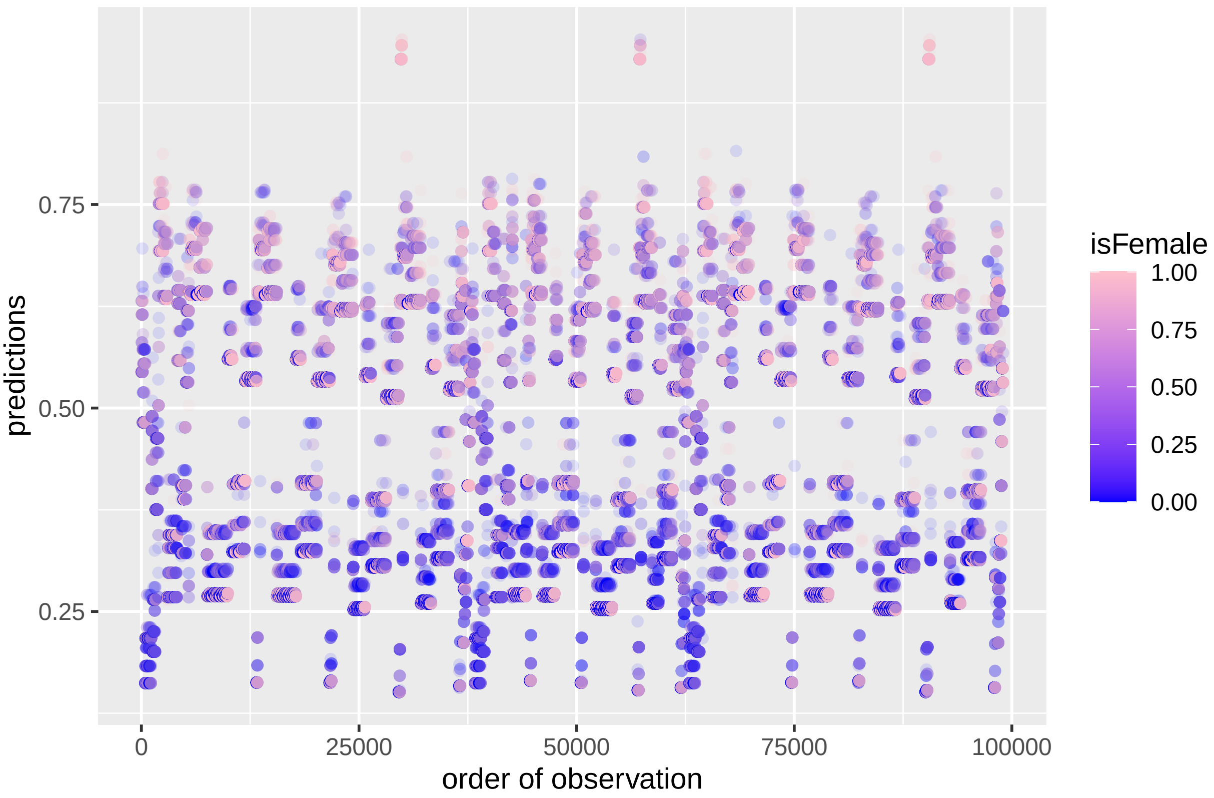
The resulting model is as follows:

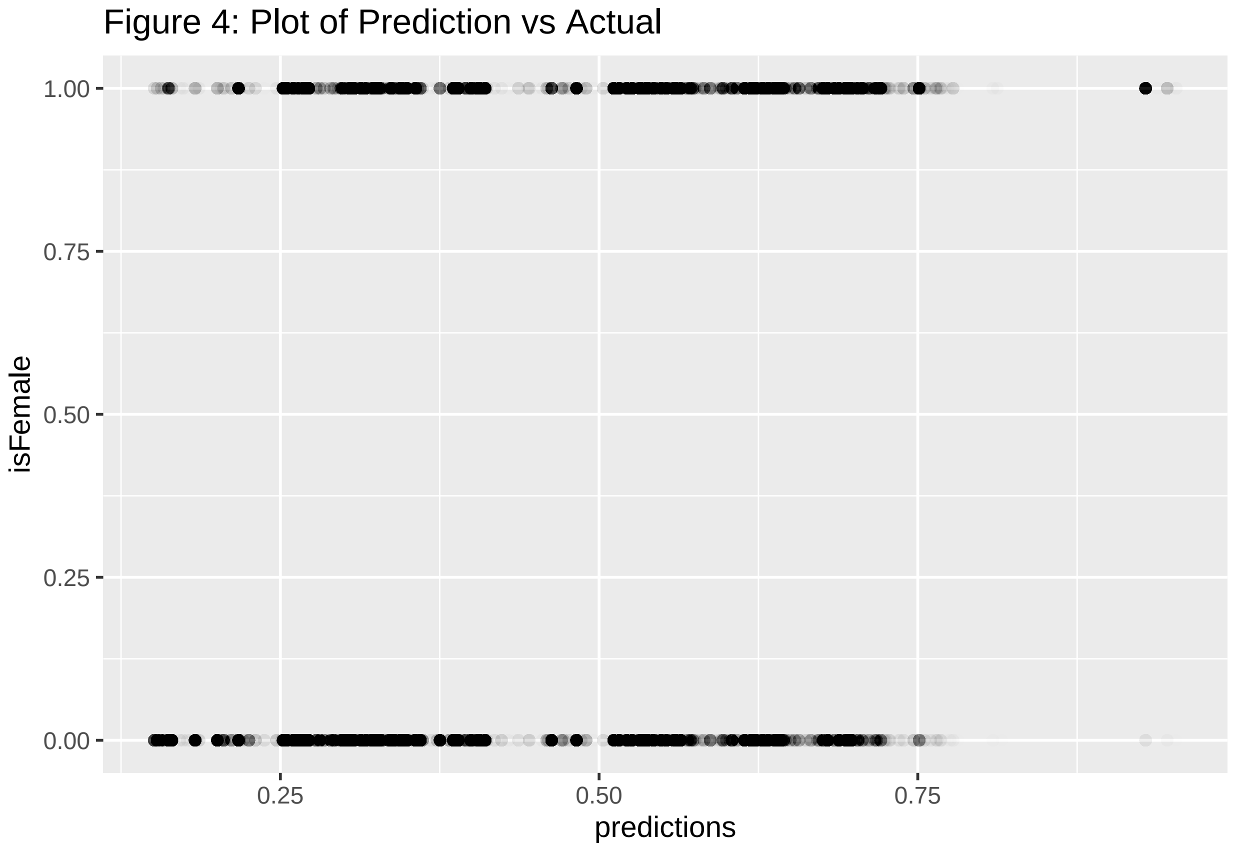
	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	0.8212	0.0302	27.19	0.0000
Study.Year2	-0.0015	0.0220	-0.07	0.9465
Study.Year3	-0.0874	0.0222	-3.94	0.0001
Study.Year4	-0.0461	0.0228	-2.02	0.0431
Study.Year5	-0.1469	0.1442	-1.02	0.3084
Study.YearD	-0.6602	0.0360	-18.33	0.0000
Study.YearM	-0.0233	0.0301	-0.78	0.4378
Study.YearN	-0.3149	0.0479	-6.58	0.0000
Faculty.GroupARTS	-0.2487	0.0293	-8.50	0.0000
Faculty.GroupCFM	-1.5753	0.1064	-14.80	0.0000
Faculty.GroupENG	-1.8210	0.0298	-61.12	0.0000
Faculty.GroupENV	-0.5796	0.0344	-16.85	0.0000
Faculty.GroupMATH	-1.5604	0.0301	-51.90	0.0000
Faculty.GroupREN	1.7752	0.2277	7.80	0.0000
Faculty.GroupSCI	-0.6878	0.0299	-23.04	0.0000
Faculty.GroupSE	-2.4593	0.0698	-35.22	0.0000
Faculty.GroupTHL	-0.8018	0.2313	-3.47	0.0005
Coop.Attendance.WorkTermCo-op Full-Time Work Term	0.0168	0.0193	0.87	0.3832
Coop.Attendance.WorkTermCo-op Part-Time Academic Term	0.2001	0.0752	2.66	0.0078
Coop.Attendance.WorkTermRegular Full-Time Academic Term	0.0169	0.0188	0.90	0.3703
Coop.Attendance.WorkTermRegular Part-Time Academic Term	0.3067	0.0310	9.90	0.0000
Visa.StatusCanadian Permanent Resident	0.1475	0.0281	5.25	0.0000
Visa.StatusInternational	0.3609	0.0192	18.84	0.0000

Table 3: Results of Multivariable Logistic Regression

Table 3 displays the summary of the resulting model, where it can be seen that the majority of variables are significant. The intercept of 0.821199 means that a reference level student receives a predicted probability of 0.6945 of being female, and each variable's coefficient is reported relative to this reference level. Relative to those in year 1, there appears to be a decreasing trend as the year of study increases, with a significant negative coefficient for the doctorate year. The reference level for faculty group is AHS (applied health sciences), and it can be seen that every faculty group except REN (Renison) has a negative coefficient. For context, Renison is an affiliated university college of the University of Waterloo and offers programs in social development studies, social work, language, culture, and arts[3]. For co-op/attendance/work-term, there is a negligible difference between co-op or workterm status, but there is a notable increase in log-odds for part-time attendance. It also appears that Canadian permanent residents are more likely to be female than Canadian citizens, and international students even more so. Figure 3 shows the predicted probabilities plotted against the order of observations of the data set with color representing the true gender, where it can be seen that pink points on average have a higher predicted probability than blue points. Figure 4 shows the true gender plotted against the predicted probabilities of the data set, where the same can be seen.

Figure 3: Plot of Predicted Probabilities





Discussion

Summary

Using the 2016/2017 University of Waterloo student data, a multivariable logistic regression model was created that predicts whether a student is female given their year of study, faculty group, co-op/attendance/work-term status, and visa status. It was found that the probability of a female student decreased as year of study increased, with graduate years having a strong negative correlation. Among faculty groups, the most positively correlated were Renison, applied health sciences, and arts, whereas the most negatively correlated were software engineering, engineering, computing and financial management, and math. Regarding co-op/attendance/work-term status, the results show no large difference for different co-op and work-term statuses, but women were more likely to be part-time. Additionally, Canadian permanent residents were more likely to be female than Canadian citizens, and international students even more so.

Conclusions

The result of this analysis corroborates many of the preconceptions surrounding gender disparity in post-secondary institutions. There may be a variety of reasons as to why there is a negative correlation with study year - it may be the case that women drop out at a higher rate, or it may be the case that the number of women enrolling to Waterloo has increased in recent years. On the topic of faculty groups, it is common knowledge that women are underrepresented in STEM and overrepresented in the humanities[4], which was also shown through this analysis. Using this model, a first-year regular full-time academic student with

Canadian citizenship would receive a prediction to be female of 93.06% if they were in the Renison faculty, compared to 16.27% for the faculty of software engineering, which gives insight into the underlying patterns within the data. There is no intuitive explanation as to the differences observed for co-op/attendance/work-term status and visa status, but they may be a result of one or many factors that affect men and women differently. All in all, it can be seen that there are a number of factors that are correlated with a student's gender.

Weaknesses & Next Steps

One major weakness of this analysis is the fact that the multivariable logistic model must exclude the first level of each category to avoid singularity, which means that the resulting coefficients must be interpreted relative to the reference level. As a next step, a more intensive analysis could be done within each category to explore differences between gender in a way that can be interpreted absolutely. As a next step, a similar analysis could be done on post-secondary students within the entire province, or on all post-secondary students in Canada, to examine how the situation in the University of Waterloo is similar to or differ from the situation on a wider scale. Similarly, the analysis could be done for past fiscal years to determine how these effects have changed over time.

References

[1] University of Waterloo. (2020). *Student Headcounts* [Data table]. Retrieved from https://uwaterloo.ca/institutional-analysis-planning/university-data-and-statistics/student-data/student-headcounts

[2] https://www.frontiersin.org/articles/10.3389/feduc.2019.00060/full Makarova, E., Aeschlimann, B., & Herzog, W. (2019, June 11). *The Gender Gap in STEM Fields: The Impact of the Gender Stereotype of Math and Science on Secondary Students' Career Aspirations*. Retrieved from https://www.frontiersin.org/articles/10.3389/feduc.2019.00060/full

[3] Renison University College. (2017, November 02). *About Us*. Retrieved from https://uwaterloo.ca/renison/about-us

[4] Hill, C., Corbett, C., & St. Rose, A. (2010, February). *Why so few? Women in science, technology, engineering, and mathematics*. Retrieved from https://time.com/wp-content/uploads/2015/05/why-so-few-women-in-science-technology-engineering-and-mathematics.pdf