STA302H1 – Final Report

Danny Chen

August 21, 2021

Introduction

The purpose of this report is to study the relationship between a student's country of origin, time spent studying for STA302H1 (weeks 1 - 4), time spent thinking about COVID-19 (weeks 1 - 4) , and their interim STA302H1 quiz scores (quizzes 1 - 3) versus final STA302H1 quiz scores (quiz 4).

Existing studies tend to focus on individual factors that affect course performance, such as the number of hours slept only, or the number of hours spent studying for a course only. However, my paper intends to explore multiple covariates simultaneously to assess their collective effect on final quiz grades, as well as the effects of two covariates on each other.

Information about Our Popoulation

Our population of interest is UofT's summer 2021 STA302H1S (July - August) cohort, which originally had 227 students at the start of the term, and 198 students enrolled as of August 13, 2021.

Experiment Description

The professor announced at the beginning of the term and in the syllabus that she would survey students on Quercus, and collect information about their quiz scores at the end of each week for the first 4 weeks of STA302H1. The first four weeks occur during these dates:

- End of Week 1 (July 5 – July 9)
- End of Week 2 (July 12 – July 16)
- End of Week 3 (July 19 – July 23)
- End of Week 4 (July 26 – July 30)

After 4 weeks, students received access to the anonymous STA302H1 performance dataset to develop a model for analysis during the STA302H1 final project.

Purpose of Developing A Model

Developing this model primarily benefits professors and students. Current professors can identify possible weak topics by identifying topics that yield the lowest quiz scores, reflect on things they did/did not help students, and then devote resources to improving lectures or creating carefully curated tutorials that address topics that students find challenging. Teaching stream professors and future STA302H1 professors would inherit these resources so they can establish reasonable STA302H1 learning goals, thoroughly prepare for more formative lectures, and address common student conceptual pitfalls that undermine student quiz

scores. When current STA302H1 students can quickly understand which factors really contribute to a high final STA302H1 grade, they have more cognitive resources available to focus on key material to getting high grades on hard quizzes and adapt to the pace of STA302H1, and they have time to implement their current study strategies or improve flawed ones in time for final assessments. Moreover, future students can establish reasonable expectations about workload and develop strategies to maximize their time and success in STA302H1 with available resources.

Plan for Developing Model

The dataset contained a small number of typos, so I opted to clean my data manually rather than programmatically. This included removing the word "hours" to safely cast numeric parts of strings as integers, removing non-Unicode characters like "", and captializing "canada" and "china", so that they would be treated the same as the countries "Canada" and "China." To finish off the data cleaning process, I decided to group similar columns (i.e., COVID times, study times, and quiz scores) together.

Some entries in the dataset contained missing (NA) data, although I deem missing quiz grades as more serious than missing countries of origin, or even missing number of COVID hours or number of STA302H1 study hours.

I could attribute missing countries of origin, missing number of COVID hours or number of STA302H1 study hours to students either forgetting, or abstaining from, share these pieces of information, while acknowledging that such students may still choose to write STA302H1 quizzes. For the purposes of preserving as much of the original dataset as possible, I decide to categorize NA countries as unknown, and leave the NA COVID and STA302H1 hours alone.

Students may occasionally miss 1 - 2 quizzes by accident due to incompatible timezones with Toronto, or because they have recently exited the STA302H1 waitlist. The best "3 out of 5" quiz marking scheme is designed to accommodate these students.

However, students who miss 3 or more quizzes jeopardize their chances of receiving a 4.0 in STA302H1 since STA302H1 quizzes weigh a total of 40% of their final STA302H1 grade. Such students may fall too far behind in STA302H1 lectures to catch up in time for quizzes. With the accelerated pace of summer STA302H1, it is much easier to fall behind and much harder to catch up if one does not commit to spending enough time with STA302H1. Unless they are experiencing extenuating circumstances that warrant a petition for additional missed quizzes, chronic STA302H1 quiz absentees are highly likely to drop STA302H1 because they have to obtain extremely high grades to compensate for missed quizzes earlier in the semester.

I will first exclude dropped students from my final dataset, since they likely do not contribute to available quiz 4 scores. I will also identify any influential outliers to remove, as no amount of variable transformations or variable re-centering can effectively correct them.

Then, I will create descriptive statistics such as histograms, boxplots, 5-number summaries, and pairs scatterplots to reveal useful relationships that will help me determine a reasonably informed, yet simple model. I will also consult scholarly articles and journals, and empirical research to propose new information to add to my model. Lastly, I will use (TODO: insert methods here) to verify my model's soundness.

Explanatory Data Analysis

There are a total of 13 variables in the dataset: a student's quiz 4 score is the response variable, and the remaining 12 variables – time spent thinking about COVID-19 during weeks 1 - 4, time spent studying for STA302H1 during weeks 1 - 4, a student's country of origin, and their Quiz 1 - 3 scores – are the predictor variables.

- Variable, Meaning, Type of Variable
- Quiz_1_Score, Student's quiz 1 score out of 10, Continuous numeric
- Quiz_2_Score, Student's quiz 2 score out of 10, Continuous numeric
- Quiz_3_Score, Student's quiz 3 score out of 10, Continuous numeric
- Quiz_4_Score, Student's quiz 4 score out of 10, Continuous numeric
- COVID..hours.W1, Time student spent thinking about COVID-19 during Week 1 in hours, Continuous numeric
- COVID..hours.W2, Time student spent thinking about COVID-19 during Week 2 in hours, Continuous numeric
- COVID..hours.W3, Time student spent thinking about COVID-19 during Week 3 in hours, Continuous numeric
- COVID..hours.W4, Time student spent thinking about COVID-19 during Week 4 in hours, Continuous numeric
- STA302..hours.W1, Time student spent studying for STA302H1 during Week 1 (can include lecture time, review time, quiz time, or assignment time), Continuous numeric
- STA302..hours.W2, Time student spent studying for STA302H1 during Week 2 (can include lecture time, review time, quiz time, or assignment time), Continuous numeric
- STA302..hours.W3, Time student spent studying for STA302H1 during Week 3 (can include lecture time, review time, quiz time, or assignment time), Continuous numeric
- STA302..hours.W4, Time student spent studying for STA302H1 during Week 4 (can include lecture time, review time, quiz time, or assignment time), Continuous numeric
- Country, Student's country of origin, Categorical/nominal

Note that quiz 1 - 4 scores are out of 10; COVID and STA302H1 study hours are measured in hours.

Relevant Tables and Figures for Noteworthy Variables

TODO: Display histograms of all predictor variables against quiz 4 score
TODO: Display boxplots of all predictor variables against quiz 4 score
TODO: Display pair scatterplots of relationships between quiz 4 score and each predictor var
TODO: Describe each descriptive statistic – "this histogram/boxplot/scatterplot displays relationship between X and Y"
TODO: Don't discuss relationship results though – See figure X in appendix
TODO: Display pairs scatterplot, with any potential outliers (influential or otherwise)
TODO: Display correlation matrix, Display linear model
TODO: Consult 3 – 4 external sources to confirm your findings.

Model Development Section

Process Used to Determine Final Model

Statistical and Empirical Justifications for Model

In-Depth Diagnostics to Verify Goodness of Model

TODO: Anything else, other than residual plot and qqplot to assess goodness of fit?