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, the time they spent studying for STA302H1 (weeks 1 - 4), the time they 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 in pedagogy tend to focus on individual factors that affect course performance, such as the number of hours slept, or the number of hours spent studying for a course. 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 a group of students from the online summer 2021 (July - August) STA302H1S cohort, which originally had 227 students at the start of the term, but has 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. Each week is specified by a date range below:

- 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 Quercus access to the anonymous STA302H1 performance dataset to develop a model for analysis during the STA302H1 final project.

Purpose of Developing A Model

The purpose of developing a model is to determine what the strongest predictors for Quiz 4 grades are: interim STA302H1 quiz scores, study time, COVID contemplation time, country, or some combination of these factors.

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 "UTF-098", and capitalizing "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.

Although some entries in the dataset contained missing (NA) data, I treated missing quiz grades as more problematic than rows with only missing number of COVID hours, number of STA302H1 study hours, or even missing countries of origin. Students may forget to, or abstain from sharing countries of origin, the number of COVID hours, or number of STA302H1 study hours – yet continue to write STA302H1 quizzes. To preserve as much of the original dataset as possible, I decide to categorize NA countries as unknown, and leave NA COVID hours 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 usually drop STA302H1 because they may fall too far behind in STA302H1 lectures to catch up in time for quizzes, or are otherwise not in a good position to commit to completing STA302H1, unless they are experiencing extenuating circumstances that warrant a petition for additional missed 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.

First, I'll exclude dropped students from my final dataset since are unlikely to contribute available quiz 4 scores. I plan to identify any influential outliers to remove, as no amount of variable transformations or variable re-centering can effectively correct them.
(TODO: Add influential outlier checks later on.)
(TODO: Also remove students without Quiz 4 scores?)

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.
(TODO: Address variable transformations to reduce skewness if necessary.)
(TODO: Address variable re-centering to reduce multicollinearity if necessary)

I will use model diagnostics to verify assumptions of my final model. Lastly, I will also consult empirical research and scholarly research to confirm my findings and propose ways to improve my model.

Explanatory Data Analysis

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

The following table describes each variable, its meaning, and its type:

- Variable, Meaning, Type of Variable
- Country, Student's country of origin, Categorical/nominal
- 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

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

I begin by creating a pairs scatterplot to get an overview of relationships between all combinations of variables. Since the pairs scatterplot is symmetric along its main diagonal, I could safely omit the bottom half of my pairs scatterplot. I then decided to analyze Quiz 1 - 3 scores, Weeks 1 - 4 COVID-19 times, and Weeks 1 - 4 STA302H1 study times separately against Quiz 4 score. I only inspect the first row where the quiz 4 score was the response variable to hypothesize any relationship between one's Quiz 4 score and each predictor variable. For simplicity's sake I decided to prioritize looking for a simple model – such as a linear relationship or a quadratic relationship – over a more complex model – such as a 3rd order model (or higher), logarithmic, or nth root relationship.

Looking at the descriptive statistics from the previous section, we see that the histograms for grades are left-skewed as few students fail quizzes, and the hisotgrams for COVID-19 times and STA302H1 study times are right-skewed as few students spend a significant amount of time thinking about COVID-19 or studying for STA302H1 compared to the mean or median amounts of study and COVID times.

We find that the R^2 value of our model is (TODO: Insert R^2 value here) whereas the adjusted R^2 value is (TODO: Insert adjusted R^2 value here).

Statistical and Empirical Justifications for Model

To show that my model is linear, I'll do the following.

1. Linearity
2. Independence of Errors
3. Homoscedasticity (constant variance)
4. Normality of Error

To verify A1, the original scatterplot must exhibit a linear relationship.

(TODO: Show plot that satisfies A1.)

To verify A2, the scatterplot of residuals versus fits for all predictor variables must not have a discernible relationship.

(TODO: Show plot that satisfies A2.)

To verify A3, the scatterplot of residuals versus fits for all predictor variables must not have a "megaphone effect" or a "bowtie effect" where residuals tend to increase/decrease as fits increase.

(TODO: Show plot that satisfies A3.)

To verify A4, we could either show that all points on a QQplot follow the QQline closely or show that the histogram of residuals is approximately normal.

(TODO: Show plot that satisfies A4.)

In-Depth Diagnostics to Verify Goodness of Model

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

Conclusion

Purpose of Final Model

Interpretation of Final Model

Remaining Limitations and Problems with Model

Proposed Improvements with Model

Generalizability of Model?