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	Continuous numeric
`STA302..hours.W2`	Time student spent studying for STA302H1 during Week 2	Continuous numeric
`STA302..hours.W3`	Time student spent studying for STA302H1 during Week 3	Continuous numeric
`STA302..hours.W4`	Time student spent studying for STA302H1 during Week 4	Continuous numeric

Note that time spent studying for STA302H1 can include lecture time, review time, quiz time, or assignment time.

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: Add 5-number summaries of all predictor variables)
(TODO: Display pair scatterplots of relationships between quiz 4 score and each predictor variable)
(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)
(TODO: Display lm() output of linear model in nice table using gtsummary().)
(TODO: Point out similarities and differences between various descriptive statistics (i.e., histograms, boxplots, scatterplots)

(TODO: Comment on distribution of variables) (TODO: Comment on 5-number summary (mean, mean, IQR), outliers across all countries – analysis will be performed later to find influential outliers) (TODO: Consult 3 – 4 external sources to confirm your findings.)

Model Development Section

Process Used to Determine Final Model

I begin by identifying influential outliers in the original dataset.
(TODO: Use Cook's distance and leverage points to find influential outliers?)
(TODO: Use VIF method to find influential outliers?)
(TODO: Use DDFITS, DDBETA to find influential outliers?)
Notice that there are no influential outliers because no points outside of cook's distance in upper right and lower left quadrants of plot, so there are no points to remove.

To derive the terms in my model, I first decide to construct a correlation matrix to find the correlation between two covariates and the correlation between a covariate and a response variable. I use highly correlated (correlation ≥ 0.50 are considered) combinations of predictor variables as a heuristic for determining potentially significant terms in my model.

Next, I create 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 the first row where the quiz 4 score was the response variable to hypothesize a relationship between one's Quiz 4 score and each predictor variable.

For simplicity's sake, I wanted to stick to using a simple model such as a linear relationship or a quadratic relationship, rather than a more complex model such as a 3rd order model (or higher), a logarithmic, or even a square root relationship. I fitted the initial model using `quiz4` as the response variable; and `quiz1`, `quiz2`, `quiz3`, `covid1`, `covid2`, `covid3`, `covid4` (including 3 quadratic terms, since `covid1`, `covid2`, and `covid4` look more quadratic), 8 interaction terms (to see how consecutive COVID and study weeks change over 4 weeks), and `country` as predictor variables.

(TODO: Show lm() results of original model in nice table.)
(TODO: Show ANOVA results of original model in nice table.)
In the original model, I found that only the predictor variables `quiz3`, `I(covid1 ** 2)`, `I(covid2 ** 2)`, `I(covid1 * covid2)`, and `I(study1 * study2)` were significant at the 5% significance level. A coefficient of 0.477087 for `quiz3` shows that for a unit increase of quiz 3 score, one's quiz 4 increases by about 0.48/10 points. A coefficient of 0.016115 for `I(covid1 ** 2)` shows that a 1 hour increase in `covid1` time increases quiz 4 scores by $(0.016115)^2$??? A coefficient of 0.023657 for `I(covid2 ** 2)` shows that a 1 hour increase in `covid2` time decreases quiz 4 scores by $(0.023657)^2$??? An interaction effect exists between `covid1` and `covid2` times, as well as between `study1` and `study2` times. A coefficient of -0.074201 shows that `covid1` and `covid2` times are inversely related, and likewise a coefficient of -0.016578 shows a inverse relationship between `covid2` and `covid3` times. The global F-statistic value is 3.098, and the global p-value is approximately 1.436×10^{-5}. The residual standard error is 1.582. The multiple R^2 value is 0.4211, which is far off from the adjusted R^2 value of 0.2851.
(TODO: See figure X below for R ANOVA output of original model in appendix)
(TODO: See figure X below for R lm() output of original model in appendix)

I chose not to perform any variable transformations on my distributions because (TODO: Why shouldn't I transform my variables in this case? Is it because the skew doesn't mean much?). I also opted not to re-center my variables since only a few entries in the correlation matrix have high correlation (correlation $\geq$ 0.50).

To refine the original model, I use backwards selection to remove insignificant terms (terms whose p-values were ≥ 0.50) and find a subset model with the lowest AIC value. I fitted the final model using quiz 4 as the response variable; quiz3, and the interaction terms `covid1 * covid2`, `covid2 * covid3`, `covid3 * covid4` as the predictor variables. I perform all further analyses on the final model.

(TODO: Show lm() results of final model in nice table.)
(TODO: Show ANOVA results of final model in nice table.)
From the final model, I decided to remove a few more terms. The `I(covid1 ** 2)` term is the only quadratic term in my final model which added a lot of complexity to my model for a negligible change in R^2 and adjusted R^2 value. Also, the `I(covid1 * covid2)` + `I(covid2 * covid3)` terms alone made it slightly more difficult to interpret the model since it was difficult to explain how COVID-19 changes between consecutive weeks, or whether it would affect the overall significance of the final model.

In the final model, only the `quiz3` predictor variable is significant at the 5% significance level. The coefficient of 0.483867 for `quiz3` suggests that for every 1 point increase in quiz 3, there is an increase in quiz 4 grades by 0.48/10 points. The global F-statistic value increased to 15.76, the global p-value increased to 1.207×10^{-10}. The residual standard error improved marginally to 15.76. Even though the multiple R^2 value dropped significantly to 0.3135, the adjusted R^2 value increase slightly to 0.2936 – and in fact, they are much closer together than in the original model.
(TODO: See figure X below for R ANOVA output of final model in appendix)
(TODO: See figure X below for R lm() output of final model in appendix)

Statistical and Empirical Justifications for Model

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

Assumption 1. Linearity

To verify A1, the scatterplot of the final model must exhibit a linear relationship.

(TODO: Show scatterplot of final model, with line of best fit.)

Since the final model contains only 1st order terms, we can conclude that our final model has a linear relationship.

Assumption 2. Independence of Errors

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

The scatterplot of residuals versus fits for `quiz3` shows no pattern.

(TODO: Show plot of residual vs. fitted quiz3)

The scatterplot of residuals versus fits for `I(study1 * study2` shows no pattern, ignoring all outliers.

(TODO: Show plot of residual vs. fitted I(study1 * study2))

The scatterplot of residuals versus fits for `I(study2 * study3` shows no pattern, ignoring all outliers.

(TODO: Show plot of residual vs. fitted I(study2 * study3))

The scatterplot of residuals versus fits for `I(study3 * study4` shows no pattern, ignoring all outliers.

(TODO: Show plot of residual vs. fitted I(study3 * study4))

Therefore, the errors are independent.

Assumption 3. Homoscedasticity (constant variance)

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 of residual vs. fitted for full model?)

Therefore, the error terms have constant variance.

Assumption 4. Normality of Error

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.

Notice that most of the points in the middle follow the QQplot line, and the left-tail points deviate slightly more from the QQplot line than the right-tail points, suggesting that our distribution of residuals is slightly more left-skeved than right skewed.

(TODO: Insert QQplot of final model.)

Moreover, the histograms of residuals for the final model looks approximately normal since the mean looks very close to 0.

(TODO: Insert histogram of residuals for final model.)

Therefore, the error terms are approximately normal.

(TODO: Insert R code to produce residual plots in appendix.)

Hence, our model satisfies the assumptions for a linear model.

In-Depth Diagnostics to Verify Goodness of Model

To validate my linear model, I decided to use a 55/45 training-testing strategy, as well as perform a t-test for significance.

55/45 Training/Testing Split

Out of $n = 142$ data points, I want to use 55% of the data points to train my linear model, and the remaining 45% to try to predict new values and create a new distribution to see if the mean of this new distribution is close to 0.

I found that the mean of this new distribution is 8.124×10^{-15} (see figure X in appendix).

(TODO: Add histogram of residuals for trained model.)

Moreover, the histogram of residuals looks approximately symmetric, with the mean of this new distribution being -0.2102 (see figure X in appendix). To be sure, since there are $n = 142$ points, the central limit theorem states that the sample mean is approximately normally distributed.

T-test for significance

For our t-test, we hypothesized that the mean of residuals was identically equal to 0. More formally,

- $H_0 : \mu_{residuals} = 0$
- $H_1 : \mu_{residuals} \neq 0$

The t-test results showed that the p-value is 0.7938, which means that we fail to reject H_0, meaning that $\mu_{residuals} = 0$ holds. The 95% confidence interval for the mean of residuals is $(-0.4510668, 0.3462948)$, and since $0 \in (-0.4510668, 0.3462948)$, we again fail to reject H_0 and conclude that $\mu_{residuals} = 0$.

Therefore, we've shown that our linear model is a reasonable model.

Conclusion

Purpose of Final Model

Interpretation of Final Model

Remaining Limitations and Problems with Model

Proposed Improvements with Model

Generalizability of Model?