**STA302/1001 Methods of Data Analysis 1**

**Video Project – due November 14, 2021 at 11:59PM ET on Quercus**

**Deadlines:**

1. Sign up your group of two by completing this form (available at: <https://forms.office.com/r/GVWSyCPED8> ) by **November 9 at 11:59PM ET.**
   * If you are struggling to find someone to work with, use the Quercus discussion board thread entitled “Finding a Video Project Partner” to post your email and local time zone, or locate someone who has already posted and is in a similar time zone with you.
     + You may choose to do the project alone however you will be taking on the entire workload of this assessment yourself. It is recommended to attempt to find a partner to help ease the workload.
   * You have until the above deadline to dissolve/change your group. However, after this deadline has passed, no changes to groups can be made.
   * **It is recommended that you find a partner as early as possible so that you maximize the amount of time you both have to work on the project.**
2. One person in the group will submit the video project on Quercus by **November 14 at 11:59PM ET**.

**Goal of the Assessment:**

This project will give you and your partner the opportunity to use the methodology in this course on a real dataset, as well as practice explaining the logic of your analysis and interpretation of results to a general audience. You will also see that it may not always be possible to obtain the perfect model by theoretical standards and that you may be required to settle for the one that is simply good enough. You will need to focus on justifying the decisions you make in your analysis to arrive at your preferred model as well as justify how it satisfies model assumptions and other properties better than other possible models. This is important since research needs to be reproducible and so justifications are very important.

**Instructions:**

For this project, you and your partner are responsible for answering the below research question based on the dataset provided. You must only use techniques taught in lectures from Modules 1 – 6 (inclusively) to answer the research question. Using techniques not from the course will result in your methods being considered as inappropriate in the rubric. You will then need to submit a 4–5-minute video presentation in which you present the results of your analysis. In your presentation, you should use slides or some other appropriate form of visual medium to present your results. How you decide to present your results (i.e. with tables/plots/etc.) is up to you but your choice should make the steps and results of your analysis clear. During your presentation, you will need to explain how you chose your preferred model and why it is better than other possibilities.

You must use mostly non-technical language in your explanation. You can assume that your audience is familiar with the general concepts of hypothesis tests, confidence intervals and p-values, but has little experience with linear regression methods aside from knowing that it relates predictor values to mean responses and that there are generally 4 assumptions that we wish to satisfy. This means, for example, if you make reference to a residual plot, you’ll need to explain why no pattern in the residual plot is a good thing. Try to avoid simply referring to formulae or theorems as an explanation, but rather focus on providing an intuitive understanding for how the methods are used to provide you the best model for the research question.

**Research Question:**

Admission rates of colleges and universities in the United States can be drastically different from one another and this can be due to a variety of reasons, such as lack of facilities, cost, or even the demographics of the applicants themselves. The purpose of this project is to investigate which of the factors/variables in the provided dataset best explains the variation observed in admission rates for 1508 colleges and universities in the U.S. While the primary use of the model you will build is for description and understanding of the factors affecting admission rates, it may also be used to predict admission rates for schools with certain characteristics. Therefore, the model you decide is best for explaining this variation in admission rates between schools must be both simple enough for various stakeholders to understand (e.g. prospective students, university administrators, etc.), but also complicated enough and with all necessary theoretical properties required to make good predictions. You will need to find a model that balances these two properties: not overly complicated or hard to interpret, but also having reasonable properties that allow for decent predictions and inference.

A data dictionary, outlining the variables included in the dataset and how they are coded/represented is provided along with these instructions.

**Submission Requirements:**

For your submission to be considered complete, please include the following on Quercus:

1. Your video presentation in one of the below accepted formats.
2. The R code used for your entire analysis (preferably in a single document).

When you create your video for submission on Quercus, you should make sure that:

* Your presentation of your analysis should be no longer than 5 minutes in duration.
* Video presentations must begin with a face-to-face introduction of all group members, with official University of Toronto student ID cards (or other valid government-issued photo ID) having been clearly presented. Members should also briefly describe their contributions to both the analysis and the presentation.
  + Failure to do so will result in a grade of 0, regardless of the quality of the project.
    - You are not required to show your faces for the remainder of the video, but it is encouraged in order to practice your presentation skills.
  + This introduction should not exceed 1 minute and will not be counted towards the allotted presentation time.
* Both members of the pair should present some portion of the analysis – it does not need to be split equally, and instead could be divided up by section. However both members should contribute to the presentation itself (separate from the introduction piece).
* Your slides/visual aids can be clearly viewed in the video and your voices can be heard throughout.
* If you use any sort of external references/resources, you should include a list of these at the end of your presentation or as a footnote on the slide in which it was used.
* Acceptable submission file formats: .mp4 media file, MS Streams link, or MyMedia link

You must also ensure that you upload your video/link to the Quercus assignment submission page NO LATER than 11:59PM EST on the due date above. There is a grace period of 1 hour past this deadline to account for longer than anticipated upload times. Any late submissions after this grace period will be given a late penalty of 10% for each 24-hour period after the deadline. This means if you submit any time in the 24 hours after the deadline (+ 1 hour grace period) you will be penalized 10%, and if you submit 24-48 hours after the deadline you will be penalized 20%, etc. This penalty is to ensure that we receive the submissions in a timely manner so that they can begin to be graded as soon as possible, and to ensure you have sufficient time to work on your next assessment.

If you or your partner experience an illness that will prevent you from completing your project and submitting on time, you may request an extension using the form for extension requests, available on the Quercus course page.

**Tips for Making Good Presentations/Slides:**

* Start early – don’t leave your whole analysis/presentation until the deadline!
* Decide what is the story you want to tell with your presentation and design your slides to focus on this story.
  + - We need to know roughly the steps you took to get to your model, but don’t need to see any R code in the slides.
* Don’t put too much information on each slide – decide what is important to highlight and make these aspects clear
* Avoid too much text which takes your audience longer to read than a picture/table
* Focus on just the important decisions you made along the way – avoid providing too much detail, instead highlighting the relevant steps that you followed
* No need to display everything – instead get creative with what you choose to show and ensure it has a purpose to the story you are telling with your presentation.
* Be sure to interpret the results you present instead of simply displaying a figure/table – talk about what is important.
* Be sure to summarize your results, conclusions and limitations of your analysis so that you highlight what worked and what didn’t.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Presentation Characteristic** | | **Excellent (3)** | **Satisfactory (2)** | **Needs Improvement (1)** | **Missing (0)** |
| **Content of Data Analysis** | **Introduction of Data and Problem** | Goal of the study was clearly introduced **AND** numerical/graphical summaries/descriptions of dataset provided **AND** interesting or relevant characteristics of data/variables highlighted with an emphasis on their possible impact on the analysis. | Goal of the study was provided but perhaps unclear **OR** numerical/graphical summaries/descriptions of dataset incomplete or not appropriate **OR** some interesting or relevant characteristics of data/variables not highlighted but relevance not emphasized. | Goal of the study not provided **OR** no summaries/descriptions of dataset provided **OR** numerical/graphical summaries/descriptions of dataset not provided OR no highlighting of interesting or relevant characteristics of data/variables provided. | No introduction provided; student went straight to the analysis. |
| **Selecting Predictor(s)** | Appropriate statistical methods used to select predictor(s) **AND** conclusions based on these methods were correct **AND** correct justifications were provided for the use of these methods **AND** a contextual reason for the reasonableness of the chosen predictor(s) provided | Statistical methods used to select predictor(s) were mostly appropriate **AND** conclusions based on appropriate methods were correct **OR** justifications provided for the use of these methods incomplete/incorrect **OR** contextual reason for reasonableness of chosen predictor(s) incomplete | Statistical methods used to select predictor(s) were not appropriate **OR** conclusions based on selection methods incorrect **OR** no justifications provided for the use of these methods or justifications were incorrect **OR** contextual reason for reasonableness of chosen predictor(s) missing. | No clear details provided about how predictor(s) selected. |
| **Model Diagnostics** | All model diagnostics have been performed and correctly interpreted **AND** issues with the chosen model(s) have been correctly identified and their impact have been explained **AND** remedies to address issues (if applicable) have been applied correctly and justified. | Some model diagnostics have been performed and correctly interpreted but some are missing or incorrect **OR** issues with chosen model(s) not correctly identified **OR** their impact has not been explained correctly **OR** remedies to address issues (if applicable) have not been applied or applied incorrectly or not justified sufficiently. | Minimal model diagnostics performed or have been performed incorrectly **AND/OR** issues with chosen model(s) not correctly identified **AND/OR** their impact has not been explained **AND/OR** remedies to address issues (if applicable) not been applied **AND/OR** without sufficient justification. | No model diagnostics performed on selected model(s). |
| **Interpretation of Final Model and Discussion of Limitations** | The final model is presented, and the relationship interpreted correctly in context **AND** sufficient justification has been provided for why this is the best model **AND** it is clear how this model meets the researcher’s intended use **AND** any lingering issues of the model and their impact sufficiently discussed. | The final model is presented, and the relationship interpreted correctly in context **AND** some justification provided for why this is the best model **OR** it is discussed how this model meets the researcher’s intended use **OR** lingering issues of the model and their impact not identified correctly or not sufficiently discussed. | The final model is presented but the relationship is not interpreted in the context of the data or interpreted incorrectly **AND/OR** insufficient justifications provided for why this is the best model **AND/OR** it is unclear how this meets the needs of the researcher **AND/OR** lingering issues not discussed or are incorrect. | No interpretation of final model provided **AND** no discussion of lingering limitations included. |
| **Overall Presentation Quality** | **Communication of Analysis** | Student spoke clearly and was easily understood and spoke at a reasonable pace (not too slow/too fast) **AND** referred to the visual aids/slides without reading from them **AND** presentation was within the time limit. | Student spoke mostly clearly and was mostly well understood **OR** delivery was occasionally rushed **OR** occasionally read off the slides or did not make sufficient reference to them **OR** presentation was at most 15 seconds over/under allotted time of 4-5 minutes. | Student did not speak clearly or was hard to understand at times **OR** delivery was very rushed, **OR** student predominantly read the slides or hardly referred to them **OR** presentation was at most 30 seconds over/under allotted time of 4-5 minutes | Presentation was over/under allotted time of 4-5 minutes by more than 30 seconds. |
| **Visual Presentation of Analysis** | Slides/visual aids were clearly organized (not too cluttered/too sparse) **AND** contained only information relevant to the analysis and presentation **AND** tables/figures are visually appealing and easily digestible in the time allotted to each slide. | Slides/visual aids were mostly well organized (a bit cluttered/sparse) **OR** contained some irrelevant information that did not benefit the presentation **OR** was missing some information that would have benefited the presentation **OR** tables/figures were a bit difficult to digest in the time allotted or were not very appealing visually. | Slides/visual aids were not well organized (too cluttered/too sparse) **AND** contained much irrelevant information that did not benefit the presentation **AND/OR** was missing crucial information that would have benefited the presentation **AND** tables/figures were very difficult to digest in the time allotted and not visually appealing. | Slides/visual aids were not used OR were not used appropriately. |