(Draft) Rubric for 432 Project 1 Proposals

Thomas E. Love

2021-01-28

Contents

What is the Purpose of this Document?

There are two purposes.

1. Provide the teaching assistants with detailed instructions on how to evaluate each part of the proposal.
2. Provide the students in 432 with a clearer understanding of what they need to do to get their proposal approved.

How Do We Grade The Proposals?

Each project is evaluated with regard to ten elements. Those elements are:

Is The Proposal Complete?

All elements of the proposal are submitted properly through Canvas, including the raw data (for projects that need to submit raw data), cleaned .Rds, .Rmd and .HTML result, as well as a one-page note from the non-reporting partner if the proposal comes from a team. Please check that:

- the Canvas submission includes the knitted HTML result, produced with `code_folding = show` used, and produced without anywhere using `echo = FALSE`, so that all code in the entire document can be seen or hidden at the whim of the reader, **and**
- any graphs or tables are completely legible, and not (for instance) outside the size of the page, **and**
- an attempt at each of the nine tasks outlined above is part of the HTML, **and**
- the student has used the template provided, or something equivalent that maintains the same headings to facilitate finding the materials so you have no difficulties using the file.

Note: In general, no other information should be submitted, although some students may also include some initial "rawer" form of the data set, in which case, that must be explained within the proposal as part of the explanation of the data source in Section 1.

We will not evaluate the proposal further until all proposal elements are successfully submitted.

Section 1 (and Section 9)

Award the point for Section 1 only if the student (or students) has:

- specified the source of the data thoroughly (and provided a link if the data are available online) in the reference section (Section 9),
- specified who gathered the data,
- specified when the data were gathered,
- specified the purpose for which the data were originally gathered,
- provided some context as to the sampling strategy and study design (for instance, the data might be a survey, or the result of a case-control study, or the result of a randomized clinical trial)

all while using complete English sentences.

Section 2

Award the point for Section 2 only if:

- the student provides a description of the subjects (rows) in the data, in a complete English sentence or two, and it makes sense to you.

Section 3

Award the point for Section 3 only if the student(s) have:

- provided code to read in the csv file as a tibble (usually with `read_csv`) or ingested the data in some other appropriate way that is readily replicable by the teaching assistant.
- provided code which tidies the tibble with `filter` and `select` to include only the rows (subjects) and columns (variables) that they will work with in the project, so that there is no completely extraneous information left,
- placed all cleaning and data management operations in Section 3.2, yielding at the end an appropriate and tidy tibble - this should include:
  - code to, if appropriate, create properly ordered factor versions of categorical variables.
  - code making whatever mutations are necessary to the data to identify the quantitative outcome proposed for linear regression modeling, and the binary outcome proposed for logistic regression modeling and the various predictors to be used,
  - code to ensure that the subject identifier is a character variable, as needed,
  - code to restrict the sample to no more than 1200 rows, as needed.

- Sanity checks and testing are an important part of verifying that the code does what you think it does. The student should unclude brief descriptions of whatever checks you do in this section of the proposal, in complete sentences between code chunks, but should not include any long printouts of data or summaries that are not helpful to the reader.
- Categorical variables should be collapsed to six or fewer categories, and all categorical variables included in the tibble should have at least 30 observations at each level, other than the subject identifier.
- Implausible values of variables should be set to NA.
- For the proposal, the student should not impute, at all. That will come when we model.

Section 4

Award the point for Section 4 only if

- an actual listing of the tibble is provided, and
- a sentence is provided that specifies the number of rows (observations) and the number of columns (variables) in the data, that accurately reflects what the tibble listing specifies, and
- there are a minimum of 100 and a maximum of 1200 rows in the tibble, and
- there are a minimum of 7 and a maximum of 18 columns in the tibble, and
- the left-most column in the tibble is a subject id code, which appears as a character variable, and there is a clear demonstration that each row has a unique identifier, and
- the tibble is saved to an .Rds file that is part of the project submission.

Notes:

- A listing of the entire data set is not acceptable. This must be a tibble listing which includes the first 10 rows, only.
- Students can check that their `subject_id` values are unique by ensuring that `nrow(tibblename)` is equal to `n_distinct(tibblename$subject_id)`.
- The minimum size is 7 columns because that counts a subject ID code, a binary outcome, a quantitative outcome, and four predictors, which is the minimum size possible to complete the project.
- The reason that the maximum size is 18 columns is that the maximum number of potential predictors is 15 (and that's only for a data set with the full 1200 observations), to which we add a subject identifying code, and a binary outcome, and a quantitative outcome, making a total of 18 variables. Most students should, in fact, have considerably fewer than 18 variables in their data set.

Section 5

Award the point for Section 5 only if:

- an appropriate and nicely formatted Table containing all of the required information is provided, and clear to you,
- the table includes the name, role, type and description of each variable, all correctly identified and in agreement with the `Hmisc::describe` results in section 5.2
- each variable's description is clear to you, and
- the quantitative outcome planned for the linear regression has more than 10 distinct, ordered, values according to the `Hmisc::describe` results, and its units of measurement are specified.
- the binary outcome planned for the logistic regression has exactly 2 distinct values, and each includes at least 30 observations, according to the `Hmisc::describe` results.
- the list of planned inputs includes at least one multi-categorical (3-6 levels) variable and at least one quantitative variable

Section 6

Award the point for Section 6 only if:

- the student provides a clear question that they hope to answer with their proposed linear model that makes sense to you, and
- the student specifies the outcome, it is quantitative, and it has at least 100 observations with complete data,
- the student provides an attractively formatted and labelled histogram of the proposed outcome,
- the student comments appropriately on the shape of the outcome's distribution, and
- the student's list of candidate predictors includes:
  - at least four predictors in total, **and**
  - at least one quantitative variable **and**
  - at least one multi-categorical variable (with 3-6 categories)
  - no more than 4 + (N - 100)/100 candidate inputs, where N is the number of rows in the tibble

Section 7

Award the point for Section 7 only if:

- the student provides a clear question that they hope to answer with their proposed logistic model that makes sense to you, and
- the student specifies the outcome, it is binary, and it has at least 30 observations in each level and at least 100 complete observations overall, and
- the student's list of candidate predictors includes:
  - at least four predictors in total, **and**
  - at least one quantitative variable **and**
  - at least one multi-categorical variable (with 3-6 categories)
  - no more than 4 + (N - 100)/100 candidate inputs, where N is the number of rows in the tibble

Section 8 is complete and correct.

Either they have the affirmation as we request it in the instructions, or not.

Section 10 is complete and correct.

Either they have the session information as we request it in the instructions, or not.

Overall Assessment

- Students will receive 1 point for successful completion of each part of the proposal.
- Students receiving a grade lower than 10 will need to redo the problematic aspects of their proposal until they reach 10.
- Students will have 48 hours after the posting to Canvas of each redo request to resubmit their work addressing the stated concerns.
- The TAs will grade each proposal initially, and will also review the first revision of any proposal.

When Will Dr. Love see the proposals?

Dr. Love will review each proposal, either:

- once the TAs have awarded it a score of 10, either in the first submission or after one revision
- or when the student has submitted a second revision, having failed to score a 10 previously.

Should Dr. Love feel the need for additional clarification, or wish to caution the student regarding any part of the project in anticipation of the final portfolio, he will produce additional comments for the student at that time. If he is satisfied, then he won't.

Reminder for Project Groups (rather than Individuals)

- Students working in groups of two will **each** need to submit something to Canvas.
  - One student will submit the proposal, and the other student will submit a single text document (Word is fine) which states that their partner will be submitting their joint proposal.
  - Revisions (as needed) should continue to be submitted by the student who submits the initial proposal.