# Statistics 512: Project Grading Guideline

# Project objectives

The project’s primary purposes are to:

1. provide you with experience applying the regression techniques taught in class with a real-world data set
2. provide you with practice in communicating about statistical analyses.

The second goal (communication) includes explaining the motivation for a particular analysis as well as explaining what you did, what the results were, what conclusions can be drawn from the results, and any important limitations to the conclusions.

# Expectations

A single copy of the Project Report will be turned in for each group. The report should be submitted on Blackboard by one of the group members, and should include the Group Number and names of all group members at the top of the first page.

Your report will consist of seven sections. The first five will follow the format of a standard scientific article, and the last two will be appendices containing your R code and output, respect- tively. Expectations for each section are listed below, with approximate guidelines for maximum length. These length guidelines will not be strictly enforced, but please bear in mind that efficiency is a key to effective communication. You should always aim to make your point completely and clearly, but as succinctly as possible.

1. **Introduction** (1–3 paragraphs) — because this is a class exercise, my expectations for the Introduction are minimal. However, you should aim to give me some idea why the subject of the analysis is interesting and provide any relevant background information. If you are working on an analysis of you own data and choose to include a more complete introduction, I will do my best to provide you with useful editorial feedback.

Unless you are using primary data (i.e., unpublished data belonging to one of the group members) the Introduction *must* include a citation of the original data source. Additional relevant citations for background are encouraged, but are not required unless necessary to avoid plagiarism (if you are not 100% certain what this means or do not know how to cite literature, **ask me**). See **References** for additional information.

1. **Methods** (2 pages) — The Methods should allow me to fully and completely understand what you did, the order you did it in, and why you made the choices that you made. Because it is the main focus of the class, I expect the Methods section to include more detail than would typically appear in a professional report or publication. Specifically, it should include,
   1. A brief description of the data, including the sampling unit, sample size, and the names, types (continuous versus discrete), and measurement units of the individual variables in the study. The response variable should be explicitly identified. If you are working with primary data, a description of the study design and data collection methods can be included if you would like feedback on them, but are not required.
   2. A description of any preliminary exploratory analyses that were conducted, the findings from these preliminary analyses, and any remedial measures that were taken as a result of these findings.
      1. This section should give me a clear understanding of what you did and why you did it. For example, “We examined histograms for each variable and bivariate scatterplots for each pair of variables. From these plots, we determined that *Y* was strongly right-skewed and had an upward curvilinear relationship with several predictors. The bivariate plots also showed that the variance around the mean trendlines was not constant. To linearize these relationships and stabilize the error variance, we applied a log transformation to *Y* in our final analysis.”
      2. If variables were excluded from the analysis at this stage (e.g., due to high correla- tions with other predictors), then this should be explained and justified.
      3. Please **do not** include large numbers of exploratory graphs in your main report. If you feel it is absolutely needed (e.g., to help explain the relationships among multiple predictor variables), a correlation matrix or scatterplot matrix (not both) may be included. In most cases, you should include these in **Appendix B** instead, and then refer to them in your main report to justify your decisions.
   3. If applicable, provide a description of the model building process that you used. This should include a description of the stages of your analysis, if there was more than one. It should also include a description of any model selection techniques that you used in each stage, including both the algorithm (e.g., stepwise selection, best subsets selection, or partial *F* -tests), the set of candidate models considered, and the criteria used to evaluate those models. How did you choose the model that you finally used for inference?
   4. A brief description of the diagnostic methods used to check model assumptions, and a description of any identified problems and remedial measures that were applied. This section should also make it clear which models were checked (if you are doing model selection on a large number of models, you may only run intensive diagnostics on a few best performing models). Again, diagnostic plots should not be included in the main report, but should be included in **Appendix B** and then referenced in the report.
   5. A description of the inferential methods used to reach conclusions about your research question(s). Depending on your goals, this may be a very simple statement (“We used a *t*-test based on the final regression model to test the null hypothesis that ice is not colder than water.”), or it may be more complex. If you use hypothesis tests, you *must* specify a value for *α*, and the null hypothesis must be clearly identified; if you you use confidence or prediction intervals, you *must* specify the confidence level. If applicable, you should also identify any multiple-testing corrections that you used and provide the corrected *α* value (multiple-testing corrections include the Bonferroni and Scheff´e corrections, Working-Hotelling confidence bands, etc.). The inferential methods that you select should be appropriate for the questions you are trying to answer.
2. **Results** (2 pages) — The Results should provide a brief, factual summary of your findings, highlighting any specific results that have special relevance to your research question. In addition to the text, the Results section should include the following items:
   1. If you conducted all or best-subsets model selection, include a table that shows the top models, selected and ordered according to whatever selection criteria you used. This table should include all models that could plausibly be considered as contenders for the

“best” model, especially if selection criteria disagree with each other. If the total number of models is small (*≤* 10), you can include all of them in this table.

* 1. An ANOVA table for your final model.
  2. A Parameter Estimates table for your final model. This table must include the point estimate, standard error, *t*-statistic, *p*-value, and confidence intervals for each parameter. Type I and II sums of squares or squared partial correlations should only be included if you are using them to help interpret the final model and draw conclusions (if you used them to diagnose multicollinearity but do not use them for interpretation, include them in a copy of the table in **Appendix B**).
  3. If they are needed to answer your research question, you may include graphical or tabular summaries of the fitted values, confidence intervals for the mean response, or prediction intervals. These can also be included in **Appendix B** and referred to in the text.

1. **Discussion** (1 page) — As with the **Introduction**, my expectations for the Discussion are minimal. It should briefly interpret your results in the context of your research question, should explain the conclusions that can be drawn about the study population and the rele- vance or usefulness of the findings in the context of any larger, big-picture topics that were mentioned in the **Introduction**. The Discussion should also address any serious limitations or caveats that apply to the model or the dataset (are there situations where the model might provide misleading or incorrect predictions?). Optionally, you may use the Discussion to com- pare your results with those from other analyses, either on similar topics or using the specific dataset that you have analyzed (e.g., if the original source used a different analytic methods, how do your results compare?). Again, if you use primary data and choose to include a more complete Discussion, I will attempt to provide useful feedback.

Please refer to the **References** section for guidelines on references and citations.

1. **References** (1 page) — You should include references to support any factual statements made in your report that are not directly based on the data and analysis that you are per- forming, unless the statement can be reasonably considered to be common knowledge (this is a subjective decision and I will be very lenient in evaluating it). In addition, you *must* include citations for any statement that quotes or paraphrases another published work. Failure to do so constitutes plagiarism.

All works cited in the text must be listed in the **References** section, and each work should be listed only once. Do not use footnotes or end notes. You are free to use whatever format you like for both the in-text citations and the full references in the bibliography, as long as the full references include author names (last/family name and initials are sufficient), a publication year, a title, and volume and page numbers (for journal articles) or book title, editor names (if appropriate), and publisher information (for books).

With the exception of data repositories (e.g., the UCI Machine Learning Repository), websites are only acceptable as references if they possess a registered digital object identifier (DOI), and the full DOI for the website must be included in the bibliographic information, along with the website URL. The repository from which you obtained your data may be cited without a DOI if one does not exist. Note that the repository is not the original source of the data.

1. **Appendix A: Code** — Appendix A should provide a complete, organized R program that generates all of the plots, diagnostics, models, and outputs referenced in the report. It should be sufficiently commented to make it easy to find relevant parts of the code.
2. **Appendix B: Output** — Appendix B should contain **selected** R output needed to justify the decisions described in the **Methods**, along with any supplemental results that are referenced elsewhere in the report. Screenshots of the raw output are sufficient in Appendix B, but they should be clearly labeled so they are easy to identify.

# Grading criteria

Grading for the report will follow the rubric outlined on pages 5 and 6. Depending on the nature of the project, evaluations will be evaluated on 14–17 specific criteria, corresponding to the major sections detailed in the **Expectations**, plus grammar and organization. Each criterion will be scored on a five point scale. Then, these scores will be summed and divided by the total available points to yield the grade for the group portion of the project. Projects that do not involve model selection will not include the model selection sections of the Methods and Results, and will have a smaller total available score.

# Peer evaluations

In addition to the Project Report, each group member should individually complete and submit a copy of the peer evaluation form, which may be found on Blackboard. The form is in and Excel file. Please download it, follow the instructions, save it as an Excel file, and submit the completed form over Blackboard. Peer evaluations will account for 25% of the project grade. **If you do not**

**complete the peer evaluation, you will receive a zero for this portion of the project grade, regardless of the evaluations submitted by your group members**.

5

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Criterion** | **0** | **1** | **2** | **3** | **4** |
| **Grammar** | Numerous spelling errors, incomplete or run-on sentences, punctuation errors, etc. throughout document | Several grammatical errors per page | A few significant grammatical errors throughout document | Very few, mostly minor errors (e.g., occasional punctuation mistakes) | No major errors and few typos |
| **Organization** | No attempt to follow organization specified in assignment | *→* | Sections organized as specified but writing within sections is disjointed or hard to follow | *→* | Paper is well organized and follows a clear, logical flow |
| **Introduction** | Section not present | *→* | Missing important background or data source | *→* | Establishes interest, gives background needed to understand data and research questions, identifies data source |
| **Methods** |  |  |  |  |  |
| Data | No data description | *→* | Information missing (see **Expectations**) | *→* | Clear, complete description |
| Preliminary analyses | Not done or no description | *→* | Missing important checks or justifying information, or includes excessive output in report | *→* | Clear, complete description that explains and justifies decisions |
| Model building*∗* | Necessary but not done or no description | *→* | Missing important information, or includes excessive output | *→* | Clear, complete description that explains and justifies decisions |
| Diagnostics | Not done or no description | Fails to check multiple assumptions | Missing important information, includes excessive output,  or fails to check some assumptions | *→* | Clear, complete description that explains and justifies decisions |
| Inference | Not done or no description | *→* | Missing important information, or  methods not appropriate for question | *→* | Clear, complete, description that uses appropriate methods |

6

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Criterion** | **0** | **1** | **2** | **3** | **4** |
| **Results** |  |  |  |  |  |
| Narrative | No narrative description of results | *→* | Missing important information, or excessive, irrelevant information | *→* | Clear, complete, description that uses appropriate methods |
| ANOVA Table | Absent | *→* | Unedited output with excess or missing information | *→* | Formatted table with title |
| Parameter Table | Absent | *→* | Unedited output with excess or missing information | *→* | Formatted table with title |
| Model Selection Table*∗* | Necessary but absent | *→* | Unedited output with excess or missing information; table unordered | *→* | Formatted table with title |
| Additional Results*†* | Used in Discussion but absent here and in Appendix B | *→* | Unedited output or unnecessary  result | *→* | Clear format, necessary, and has title |
| **Discussion** | Absent | *→* | *→* | *→* | Interprets results and draws conclusions in context of research question |
| **References** | Absent*††* | *→* | Incomplete information; unused or inappropriate references | *→* | Complete and appropriate |
| **Appendix A** | Absent | Incomplete | Disorganized | *→* | Complete and organized |
| **Appendix B** | Absent | Incomplete | Disorganized | *→* | Complete and organized |

*∗* Model selection information is necessary only if multiple models are considered.

*†* Any additional results that you include will be evaluated on a case-by-case basis. If you do not refer to the results in your Discussion,

you should not include them in the main body of the report.

*††* Flagrant plagiarism will be considered to a violation of academic integrity. See the Syllabus for relevant policies.