Integrated Experience component of Stat 525

Stat 525 has the IE project involving analysis of real data that will be done in groups of three to five students. There are five Steps. Only the last step is done/graded individually. The report will be graded on content and formatting. Pretend the article is going into a research journal like *Science* or *Nature*, so be analytical and serious. You must use (La)TeX.

Steps 0, 1, 2 and 4 (Step 4 incorporates Steps 1 and 2) will not be individually graded; however, I will factor in how you address the feedback when I give you a group-based grade for Step 4. Step 3 will be a separately graded group-based grade. Step 5 will be a separately graded individual-based grade. Weights are: Step 3 20%, Step 4 (which incorporates Steps 1 and 2) 70%, Step 5 10%.

1. Overview of the different Steps in this project

1.1. **Step 0, due Feb 8.** I will form IE groups on Moodle by Feb 2 (2 days after add/drop-deadline). Use the Moodle group-specific forum to find out who your groupmates are and what is the best way to communicate. By Feb 8, one person from each group should send me an email discussing which project you plan to do. Make sure to cc your members so that I know you have reached consensus on this. At most two groups can share a common project, and it is first-come-first-serve. The group choices will be posted on Moodle. Make sure to have a back-up choice in case you don't email me in time. My advice is to choose something that interests you. You don't have to be a future NBA manager, real estate mogul or cancer clinician to take an interest in doing the report on NBA salaries, Oregon housing prices, or cancer malignancies. However, some background knowledge and/or a willingness to do some background research will go a long way to defining your goals of the projects, formulating meaningful hypotheses, and usefully interpreting your results.

1.2. **Step 1, due Feb 23.** On Moodle submit a report of 1.5 to 2 pages in LATeX (specifications on the fontsize is at the end). One submission per group. This should address the following.

- General background of study. Cite some literature. For some projects pre-existing literature may not come naturally, so you can stretch the relevancy. For example, if you use the data Oregon housing prices from several decades ago, maybe you can cite studies of current real estate pricing.
- Main goals of the project. What do you hope to explore? Do you have any hypotheses? Since you will have just started Stats 525, I expect these could change in the final report as you have a better understanding of what the course is about.
- Data. Briefly describe the variables ("10 variables are on this, 5 variables are on that"). What's the sample size? What do the samples represent? Where did the data come from?

If one of your groupmates is slow to respond and creating more work for you instead of helping you do yours, please let me know. This needs to be an email coming from the rest of the members so that I know there is consensus on this. **I will remove the problematic member from your group.** After this, group membership will not change. While allowances are made if you are sick or facing a sudden hardship, you cannot rely on your groupmates to carry your workload. It might be best for you to withdraw from the class and take it later if your condition has a long-term negative impact.

1.3. **Step 2, March TBD.** Initial exploring of data and single-variable regressions will be completed. This should be 4-5 pages. Each group must have a Zoom meeting with me for feedback on this, so email me (as a group) the times that all of you can meet in the week after you turn this in.

- In all the projects the ultimate goal is to build a linear regression model with some or all of the predictors available to you. In all cases you will eventually do some model building as you may not need to use all variables (or it may not be wise to use all variables). It might also be necessary to transform some of the Xs (independent variables).
- This step in the project is to explore the data in some detail in advance of fitting multiple linear regression models. Primarily we want to get a sense of the data. More specifically, what you should do is:
  - Construct a boxplot/histogram for Y and each X_i and scatter plots for each Y, X_i pair.
  - From the boxplots/histogram and scatter plots, check
    - if there are any extreme values/ potential outliers we need to worry about
    - if there is a linear relationship between Y and quantitative independent variables, X

    - if there is any pattern between Y and categorical/ qualitative independent variables, X if we might need to transform X or Y?
    - if there are X's which are highly correlated with other X's?
  - Get descriptive statistics on each of the variables (each X and the dependent variable). For example, how confident are you $\beta_1 \neq 0$?
  - Get correlations among all of the pairs of (quantitative) variables. You will want to get Pearson correlation.
  - For each of the predictor variables that are categorical/ qualitative, describe the pattern of Y over each of the possible categories of X.
  - Fit a simple linear regression model for Y and each X, and check if the assumptions of the simple linear regression are acceptable. If some assumptions are violated, consider alternative ways to resolve the issues such as transformations of Y and X and detection of potential outliers.

This will become part of the final report, with changes as needed.

1.4. **Step 3, due April 24.** Pre-record a presentation and upload to Google (site TBD). This will be shown during the last week of class, where your group will answer questions from the other students.

- The group will work on the presentation jointly. Each member should talk roughly the same amount of time.
- Each presentation will last 15 minutes and a short (1 or 2 minutes) discussion will follow.
- The presentation will include quick background, a summary of main results and conclusions.
- You will also monitor and answer any questions posed on Piazza on your project for 48 hours after the presentation of your project.

The basis of this presentation will be Step 4, and possibly parts of Step 5. So you will need to have completed most of the analysis by then even if it is not yet written it. You can incorporate Step 3 feedback to improve the Step 4 report. It's possible due to timing that you might not have included Chapter 11 remediation of influential points in Step 3 and may want to put it Step 4.

1.5. **Step 4, due during the Stats 525 final exam time.** Write your main data analysis report. This will incorporate content from Steps 1 and 2 which should take up about 4-5 pages. You may need to slim some stuff down here. In total (not counting bibliography and Step 5), the main report should be about 13-15 pages. But it can be less if it is well-written. Adding non-essential material to increase length will be penalized. Make sure to keep it as one voice. I should not be able to tell where one person ended and another started. On the one hand this should be technical and you do not need to explain in the main report stuff that your Stats 525 students also know. (Do not use methods from other classes you may have taken, like machine learning, because the point of this IE project is learn and better understand the theory of linear regressions.) On the other hand, someone without a statistical background should be able to take something away just by reading the the beginning part of the paper where you state your goals and the (individual Step 5) conclusion part of the paper.

For this main portion of your report, you need to use what you learned in Steps 1 and 2 to address your clearly stated goals. I should not have to search hard to find these goals. For most of the report, you will draw from what you learned in Chaps 6-10. For example, if you have many quantitative variables (which most projects do), maybe the report should emphasize the model selection flow chart of Chapter 9. On the other hand, if you have a smaller number of variables and a fair number of them are qualitative, maybe you should run more residual diagnostics in Chapter 7 and 8. Often times, both of these will be needed. In all cases well as there should be a discussions of outlier/influential points in Chapter 10. If you have strong evidence of influential points, is your model different when you throw out that point?

Here are some specific points you should address.

- Be clear to highlight what are the goals of this project. For example, maybe you want to focus more on building the best multivariable model (for example with Chap 9), or on the individual impacts of each each variable on the response (for example with Chap 7). Or maybe both.
- Which analyses have been run? How did you get your initial sets of models?
- In building your final model, discuss which interactive terms and higher order terms were included. How did you diagnose and remedy correlation?

- Diagnose and remedy outlier and influential data points.
- Include a statistical summary of the results.

Note that you can still get an excellent grade even if you if you cannot conclude with a definitive outcome, as long as I see that you put in a thorough amount of work, for example justifying *why* you did not reach a definitive outcome.

1.6. **Step 5, due during the Stats 525 final exam time.** Write your own individual 2-page conclusion. This is not included in the 13-15 page count in Step 4. You can spend some time reviewing what you did (but not very much). Give a subject matter interpretation of the results and the implication of those results. Draw some conclusions. What are the implications of your model. What is expected or unexpected? Which answers were definitive. Why might some of the outcomes not been definitive? Imagine this 2-page conclusion immediately follows the Step 4 group report. Try to make it seamless and don't spend time/space repeating material as if it were a stand-alone 2-page paper.

2. Details for Step 3

We will watch your 15-minute presentation in class. Roughly speaking, you should think about allocating time in the following manner:

- 4 minutes for an introduction into the nature of the data, how it was collected, and the question(s) of interest
- 8 minutes for describing your model and how you came up with it, probably accompanied by some exploratory graphs and descriptive statistics
- 3 minutes to describe your results and conclusions

Your grade will be based on the following: reasonable and appropriate choices made in analyzing the data; insightful description of the research question and conclusions; quality of presentation (interesting, easy to follow, slides and organization were clear, table and graphs were readable); answering in-person and Piazza questions.

Please keep the following questions in mind as you prepare your presentation (some questions may not apply to certain types of projects):

- What is the main question I am trying to answer?
- How were the data collected/gathered/sampled?
- Are there any confounding relationships present?
- Are there any interactions present?
- Is your model reasonable?
- What assumptions is it making?
- What are the limitations of my analysis (assumptions which may not hold, limitations of the data, etc.)?

3. Details for Steps 4 and 5

3.1. **Layout.**

- The title and authors's names should appear on page 1. Subsequent pages are numbered.
- Step 4 should be 13-15 pages. 10-12 point font with 1.5 line spacing is appreciated, and you can play with the margins to within reason.
- Graphs should not take up more than 1/2 a page, and 1/4 size graphs are fine unless some detail needs to be examined closely. Graphs should have numbers and self-descriptive titles.
- All tables should have numbers and self-descriptive titles. Regression results can be put in tables, and only use what you talk about (no residual values, for example).
- All tables, confidence intervals, and p-values should be in readable numbers, that is no scientific notation unless really necessary. Tables may have to be reformatted to meet these requirements. Hypothesis test results should be reported in line, even if the result is in a table. For example: The effect of chocolate was significant ($\beta = 12.35, t = 2.78, p = 0.032$).
- Overall, you do not need to report every graph, every output, etc. For example, I expect you to check for regression assumptions. If they fail, show why. If they pass, a 1/4 size plot of the residuals vs. fitted values and Q-Q plots should be fine.
- No R code or R table output anywhere. R code will be submitted separately (see below).

3.2. **An approximate rubric.**

- Content:
  - Excellent: Exceptionally well-presented and argued; ideas are detailed, well-developed, and supported with specific evidence and details. Appropriate use of supporting figures.
  - Good: Well-presented and argued; ideas are detailed, developed and supported with evidence and details that are mostly specific.
  - Fair/Poor: Content might be reasonable but ideas are not particularly developed or supported; some evidence, but usually of a generalized nature.
- Structure/organization:
  - Excellent: Well-planned and well-thought out. Writing shows high degree of attention to logic and reasoning of points. Includes a descriptive title, introduction, statement of main idea, and conclusion. All paragraphs have clear ideas and smooth transitions.
  - Good: Good overall organization. Most paragraphs have clear ideas and transitions.
  - Fair: There is a sense of organization. Some points are misplaced and/or stray from the topic. Some paragraphs have clear ideas; transitions are weak.
  - Poor: No sense of organization.
- Style:
  - Excellent: Sentences are clear and concise. Speaks with a single voice and in a consistent manner. Impossible to tell who wrote what.
  - Good: A few awkward sentences and run-ons. A few inconsistencies in style. Single voice.
  - Fair: Mutliple awkward or unclear sentences. Can determine where one person's contribution ends and the next's start although a respectable transition still occurs.
  - Poor: Prevalence of unclear sentences. Parts may be painful and difficult to read. Looks like several short reports collated together.
- Formatting and grammar:
  - Excellent: Margins, spacing and indentations are correct. Correct formatting of sections and equations. Figures are captioned and labeled. Perfect grammar, spelling, syntax and punctuation.
  - Good/Fair: A few minor errors in formatting, or there are areas where formatting could be improved. A few errors in grammar, spelling, syntax and punctuation, but not many.
  - Poor: Consistent errors in formatting. Shows a pattern of errors in spelling, grammar, syntax and/or punctuation. Little to no proof-reading.

3.3. **Additional comments.**

- Your conclusion/results section should describe the model results in the context of the problem. Be sure to explain the meaning of interactions and inclusion of categorical variables. If you have multiple models, discuss the differences in terms of the model. Again, a non-statistician should be able to understand your final model by reading your conclusion.
- Don't forget I would like a summary of the data in the introduction section.
- Please upload your R code and any data files you use. I want to be able to reproduce your results. This is separate from the report and not counted in the page length. There should be no R code in the main report. Tables of data and parameter estimates/output are useful, but again, don't just paste in R output.
- When it comes to model selection, no need to detail which variables were dropped. Just say, for example, "Backwards stepwise selection was used on the model with all second order interactions". If your starting model is non standard, say you eliminated interactions with gender, specify that model explicitly.
- Don't forget to back transform.
- Make sure to discuss assumptions. Proof, or lack thereof, can be given by the results of a hypothesis test (test stat= $12.34, p < .001$) or a graph. Please, no Box-Cox graphs, just tell me the results. Residual plots can be useful.
- Justify any removed points.