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. Note that only the last part is done 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. I strongly recommend using LaTeX.

1. Timetable

1.1. **Step 0, due Feb 16.**
Send me an email discussing which project you plan to do. At most two groups can share a common project, and it is first-come, first serve.

1.2. **Step 1, due Feb 23.**
A report of at most 2 pages (all pages are single-spacing and single-sided), common for the group.

- 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. Do you have any hypotheses?
- 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.

Part of this, like citations, will go into the final report.

1.3. **Step 2, mid March TBD.**
Initial exploring of data will be completed and the summary report. There is a page limit of 5 pages. Each group must have a meeting with me.

- 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 X, and construct scatter plot(s) between Y and all X's
  - From the boxplots/histogram and scatter plot(s), 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 a Y or an X or consider adding a function of X (e.g, X^2)?
    - 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)
  - 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 regres- sion 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 may become part of the final report, with changes as needed.

1.4. **Step 3, last week of class.**
Pre-record a presentation. 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.

1.5. **Step 4, due last day of finals.** Main report. There is a page limit of 10 pages. The report should follow the model selection flow chart of Chapter 9.

- Which analyses have been run. How did you get your initial model? Discuss which interactive terms and higher order terms were included.
- A statistical summary of the results.
- A subject matter interpretation of the results and the implication of those results.

This can be technical.

1.6. **Step 5, due last day of finals.** Write your own individual 2-page conclusion. You can spend some time reviewing what you did. But 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 10-page group report. Try to make it seamless and don't spend time/space repeating too much as if it were a stand-alone 2-page paper.

2. Expectations for Group Presentation

You will give a 15-minute presentation on your data and your analysis of it. These presentations will take place in our classroom during our class time. 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

You will have 15 minutes to speak; plan accordingly. Have your group do several practice versions beforehand. There will be a few minutes (1 or 2 minutes) for questions after each talk. Your grade will be based on the following:

- 30% Reasonable and appropriate choices made in analyzing the data
- 25% Insightful description of the research question and conclusions
- 25% Quality of presentation: interesting, easy to follow, slides and organization were clear, table and graphs were readable
- 20% Answering 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. Expectations for Project Report

3.1. **Layout.**

- Include the title of the project and the names of group members in the first page.
- Put a page number in each page.
- There is a page limit of 10 pages . 10-12 point font with 1.5 line spacing is appreciated, and you can play with the margins and such to conserve paper.
- 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.

3.2. **Format.**

- Introduction. You will need an introduction that fleshes out the data set, where it came from, and what you hope to accomplish.
- Methods/Results. This will be the longest section. Describe the methods you used and important results. Again, I don't want everything, just the important stuff. It should start with a summary of the data and the go from there.
- Conclusion. Sum up the results. What is the take home message. I always like to think that a person should be able to read the introduction and conclusion and come away with some information.

3.3. **R code.** Please upload your R code and any data files you use. I want to be able to reproduce your results.

3.4. **Additional comments.**

- Please make an effort to make your final paper look "nice". In other words, if your paper looks as it has been thrown together, your grade will reflect that. Remember, I am only looking for graphs that help tell your story and you will need to email me your R code, so don't clutter the paper up with that. Tables of data and parameter estimates/output are useful, but again, don't just paste in R output.
- 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.
- Don't forget I will need your R code and any data files called by the R code.
- 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.