STAT 525 Integrative Experience (IE) Project

Maryclare Griffin

Spring 2020

1 The Integrative Experience (IE) component of Stat 525

The General Education requirements include an integrative experience (IE) component (see http://www.umass.edu/gened/objectives-designations/curricular-designations/integrative-experience). The Department of Mathematics and Statistics has several upper level courses with an IE component that meet the IE requirement and one of them is Stat 525. Stat 525 is meant to meet the IE requirement for those doing the Statistics or Actuarial concentration or in other concentrations but with statistical interests. The university's IE requirement has three criteria (from the General Education Guidelines):

1. *Providing a structured, credited context for students to reflect on and to integrate their learning and experience from the broad exposure in their General Education courses and the focus in their major.*
2. *Providing students with the opportunity to practice General Education learning objectives such as oral communication, collaboration, critical thinking and interdisciplinary perspective-taking, at a more advanced level.*
3. *Offering students a shared learning experience for applying their prior learning to new situations, challenging questions, and real-world problems.*

Stat 525 includes an IE project involving analysis of real data that will be done in groups of four to five students. The project will require:

- A careful examination/reflection of the context in which the data arose and associated background along with a discussion of why the problem is of interest in a general sense.
- A full discussion of what the results mean from the perspective of how those results may change ones thinking about the problem at hand or suggest future changes, such as a change in management policy or a suggested change in behavior. Included here should be a discussion of how the data used arose and any cautions that might be needed in how it is used to answer the original question of interest.
- Group presentation of the results to the class. There will be two classes with 4-5 projects presented in each class.
- Final project report should be individually done. This can incorporate discussion from the presentation sessions.

2 Timetable

2.1 Step 0

(Optional) Students can submit suggestions for additional project topics (see "Stat 525: Spring 2020 IE project topics" for initial project topics). A description to the data set, a link to where it can be obtained, and at least one reference for an interesting scientific question that the data could be used to analyze is due Thursday January 30, 2020. Approved datasets will be added to the list of possible project topics.

2.2 Step 1

Students will form groups of 4-5 classmates. Each group will read the project topics and their descriptions/data listed by me (see "Stat 525: Spring 2020 IE project topics"). A report of at most 1 page, common for the group, will be handed in by February 6, 2020. This should include

- The names of the group members,
- Ranked preferences for the project topics.

I will assign projects to groups based on preferences.

2.3 Step 2

A report of at most 2 pages, common for the group, will be handed in by February 20, 2020. This should include.

- Main goals of the project
- General background,
- Discussion of the problem. This will become part of the final report, with changes as needed. Each group must have a meeting with me during the office hours.

2.4 Step 3

Initial exploring of data will be completed and the summary report will submitted to Moodle by 11:59pm, April 2, 2020. There is a page limit of 5 pages (double-sided). Each group must have a meeting with me during the office hours.

- 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) and possibly transform the Y (dependent variable).
- 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 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 may become part of the final report, with changes as needed.

2.5 Step 4

In order to show the project in progress, a draft of statistical analyses will be handed in using Moodle by 11:59pm, April 21, 2020. There is a page limit of 10 pages (double-sided). Each group must have a meeting with me during the office hours. The report should include:

- Which analyses have been run
- A statistical summary of the results
- A subject matter interpretation of the results and the implication of those results

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

2.6 Step 5

Presentations in class April 23, 2020 (4 groups) and April 28, 2020 (4 groups).

- The group will work on the presentation jointly.
- 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.

2.7 Step 6

The full report is due by May 4, 2020 (the last day of final exams) at the latest and it should be done separately by each individual. This can incorporate discussion from the presentation sessions. You should be able to finish earlier but this will give you flexibility depending on your exam schedule.

- While the background and statistical analysis (including statistical conclusions) will be common to the group, the last couple of pages of the report, in which you will discuss what you have learned about the problem, must be done individually.
- There is a page limit of 12 pages (double-sided), but an organized and well- written report of 10 pages is absolutely appropriate

3 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. 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.)?

4 Expectations for Project Report

4.1 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 12 pages (double-sided). 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 (β=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.

4.2 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.

4.3 R code

Please email me your R code and any data files you use. I want to be able to reproduce your results.

4.4 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 sent to me via email.
- 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.

5 Dig In!

For me, analyzing data is the most exciting part of being a statistician! Eminent statistician John Tukey famously said:

The best thing about being a statistician is that you get to play in everybody's backyard.

Regression is a powerful tool for data from nearly every area of quantitative study. I hope you are excited about combining your study of statistics with your general knowledge and experience to gain insight into a substantive problem.