Lab 2: What Makes a Product Successful?

w203: Statistics for Data Science

Introduction

This lab is a group assignment taking place over 3 weeks. The centerpiece of the lab is an original regression study, based on a research question of your choosing. You will present your findings in a pdf report and in a short slide presentation to your classmates. The lab also includes intermediary deliverables, with something due every week until the end of class. The following table summarizes all deliverables.

Deliverable Name	Due Date	Grade Weight
HW 11 - Explanation Practice	Tues, July 19 2pm	Part of HW
Part 1 - Research Proposal	Tues, July 19 2pm	5%
Part 1 - Within Team Review	Tues, July 19 2pm	5%
HW 12 - CLM Practice	Tues, July 26 2pm	Part of HW
Part 3 - Final Report	Tues, Aug 2 2pm	85%
Part 3 - Final Presentation	Unit 14 Live Session	5%

Note that we are including Homework 11 and 12 in this lab packet. You may work with your teammates on these homeworks, but they continue to count towards the homework portion of your final grade.

This is a group assignment. Your live session instructor will coordinate the formation of groups. We encourage you to use the lab as an opportunity to learn how to work as a team of collaborating data scientists on shared code; how to clean and organize data; and how to present work in a compelling way. We hope you will create an environment in which individuals can take risks, and improve the skills they are most interested in – they might be project coordination, management of code, plotting, or others. *We hope that you can support and learn from one another.*

Do not share your work outside of class. Please keep your work private so that future students can use the same data to learn.

Some Encouragement for the Project

This project touches on many of the skills that you have developed in the course.

- When you perform a regression analysis, you are using a statistical model to represent the world. Your data actually results from physical processes: atoms bouncing around, sunlight and use weakening product components, manufacturers responding to each other's prices. Your model instead assumes that the data is drawn from probability distributions.
- This class is not a class in pure theory! Your team will have to evaluate your statistical assumptions, using your background knowledge, visualizations, and numerical summaries.
- Throughout, you will have to communicate both to a technical and non-technical audience.

Finally, have fun with this project. You have worked hard this semester to build a foundation for reasoning about the world through statistical models. This project is a chance for you and a team of peers to apply this reasoning.

Part 1 - Explanation Practice

Imagine you are working at a company that is considering whether to purchase a cybersecurity training for its employees. A colleague has gathered data about a set of Fortune 500 companies. They present the following regression estimate:

Outcome variable: Security Breaches	
Cybersecurity Training Hours	0.052** (0.009)
Emails Encrypted	-1.23*** (0.01)
Constant	10.790** (5.078)
Note: *p<0.5; **p<0.01; ***p<0.001	

Here, Emails Encrypted is a covariate that measures the fraction of employees that reported using encrypted email either "all of the time" or "most of the time" in a questionnaire. Based on this regression, your colleague suggests that cybersecurity training leads to more security breaches, so your company should not invest in it.

You are convinced that the statistical assumptions underlying the estimate are sufficiently justified. However, you have concerns with your colleague's causal interpretation. Provide a response as follows:

1. An omitted variable is whether the company has high value assets that might be attractive targets for criminals. Argue whether the omitted variable bias is towards zero or away from zero (5 sentences max).
2. Explain why there is a possibility of reverse causality. Argue whether the direction of bias is towards zero or away from zero (5 sentences max).
3. Explain why there is an outcome variable on the right hand side. Argue whether removing it would make the coefficient on Cybersecurity Training Hours move up or down (5 sentences max).
4. Provide a one-sentence conclusion addressing the idea that your company should not invest in cybersecurity training.

Part 1 - Research Proposal

Imagine that you are part of a team of product data scientists at Acme, Inc. Your manager, Mx. Coy Ote, has given you the freedom to choose your own product to investigate, and evaluate a way to make it more successful.

Your task is to select and develop a research question, find appropriate data, then write a proposal for the research study you will perform.

Research Question

Your research question must be specific, it should clearly state an X and a Y, defined at a conceptual level. Your X should be a design property or characteristic of a product that could be modified in the production process, and your Y should be a metric of success.

In selecting your research question, you will have to use the skills you developed in RDADA to work on a question that can be addressed using a regression analysis. It is not appropriate to ask "What product features increase success?" or "How does product design affect sales?". These types of questions are not amenable to a modeling based approach and your study would likely become a fishing expedition. Instead, your team will have to use your background knowledge to identify a relationship you want to measure between a specific design feature and a specific metric of success.

If your data set is large enough, you can begin your process by splitting the data into an exploration set and a confirmation set. As a rough guideline, you might put 30% of your data into the exploration set, but make sure that both sets have a minimum of 100-200 rows of data. Use the exploration set to build your intuition, explore the data, and build your model specifications. In the ideal case, all *modeling decisions* that you make are based on the exploration set. The confirmation set should be used only once the code to generate your regression table and other results is set. All numbers in your report, as well as your discussion and conclusions, should be based on your confirmation set.

Because your manager is interested in *changes* to a product, they are fundamentally asking you to perform an explanatory study. As we have noted in the class, given observational data, an OLS regression is usually not a credible way to measure a causal effect. We have purposefully selected a domain in which the one-equation structural model is at least partially defensible. The most prominent causal pathways will go in one direction, from product design characteristics to success. While not a perfect reflection of reality, we expect your model to be plausible enough to make your results interesting. At the same time, you will need to analyze potential violations of the one-equation structural model and what effect any violations may have on your results.

Data

For this lab, you and your team will be responsible for gathering the data that you use. The data should be publicly available, and should be relevant to your research question. To increase the diversity of products investigated, we are asking students to avoid working on data that is sourced from Yelp and Airbnb. Also, please do not use the Ames housing data analyzed in the sample answer. There are many public data resources available, for example:

- New York Times
- Tidy Tuesday
- ICPSR for social and political data
- Data.world
- Dataverse for published research data
- UC Irvine Machine Learning Data Repository
- Google Dataset Search
- Amazon Open Data Registry
- Azure Open Data Registry

Requirements for your data:

- Data should be cross-sectional (i.e. not have multiple measurements for a single unit of observation). You may, for example, take a single cross section from a larger panel, or combine measurements from different time periods into a single number.
- We recommend a minimum of 100 or 200 observations. A team can choose an interesting dataset that is smaller than this, however, this will then require the team to assess and satisfy the more stringent CLM assumptions.
- The outcome (or outcomes) that you use should be plausibly metric (i.e. number of sales of a product; number of views of a video). For this lab however, to make it easier to find data, teams may use an ordinal outcome variable if necessary. If using an ordinal outcome such as a 1-7 Likert scale, the team should clearly highlight this limitation in their report.
- For any omitted variable that would call your results into question, the data should include the best possible variable that operationalizes this concept. At a minimum, the data should have a variable that serves as an imperfect measure - or *proxy* - for the omitted variable.

You may draw different variables from different data sources. You may use data sources not on the above list. You must document any data source you use in your report.

Example of a Research Question

Suppose that your team is interested in learning how the length of lanyard attached to a catapult affects performance. (A classic question from Roadrunner cartoons.)

You work to develop a primary outcome: proportion of boulders that land on their target.

On Acme's servers, you find data on lanyard length, maximum-rated weight for the catapult and sales region. However, when you are reasoning about the product, you also note that length of the catapult arm and size of the catapult wheels are also likely to affect performance and are correlated with lanyard length. Because any model that does not include these confounding variables would yield estimates that conflate the importance of wheels and arms with the lanyard, you determine that the off-the-shelf data is not complete and that you need to encode the data yourself.

In the modeling phase of your project, your team proposes to build three models. One model estimates the relationship between targeting accuracy and lanyard length by itself. A second model is similar, but adds a set of covariates including length of catapult arm and size of catapult wheels. Finally, a third model includes an interaction term between lanyard length and customer type (first time or repeat), allowing you to investigate whether the effect of lanyard length is heterogeneous depending on the person operating the catapult.

Research Proposal

After a week of work, the project team will submit a research proposal. The maximum length is one page. This is so that your instructor can provide feedback to all teams quickly.

Please answer these three questions in your proposal:

1. What is the research question? Specifically, what is the X concept and what is the Y concept?
2. What is the data source? What variables will you use to operationalize X and Y?
3. What is the unit of observation? That is, does each row of the data represent a person, a review, a hotel stay, or something else?

The research proposal is intended to provide a structure for the team to have an early conversation with their instructor. It will be graded credit/no credit for completeness (i.e. a reasonable effort by the team will receive full marks). Your instructor will read these proposals and will contact the team with any necessary course corrections, suggestions, or feedback.

This proposal is due at the start of week 12, in Gradescope, with one submission for the whole team.

Within-Team Review

Being an effective, supportive team member is a crucial part of data science work. Your performance in this lab includes the role you play in supporting your teammates. This includes being responsive, creating an environment in which all members feel included, and above all treating each other with respect. In line with this perspective, we will ask each team member to write two paragraphs to their instructor about the progress they have made individually, and the team has made as a whole toward completing their report.

This self-assessment should:

- Reflect on the strengths and weaknesses of the team and the team's process to this point in the project.
  - Where your collaboration has worked well, how will you work to ensure that these successful practices continue to be employed?
  - If there are places where collaboration has been challenging, what can the team do jointly to improve?
- If there are any individual performances that deserve special recognition, please let your instructor know in this evaluation.
- If there are any individual performances that require special attention, please also let your instructor know in this evaluation.

Instructors will treat these reviews as confidential and will not take any action without first consulting you.

This reflection is due at the start of week 12, in Gradescope and requires one submission per person.

Part 2 - CLM Practice

For the following questions, your task is to evaluate the Classical Linear Model assumptions. It is not enough to say that an assumption is met or not met; instead, present evidence based on your background knowledge, visualizations, and numerical summaries.

The file `videos.txt` contains 9618 observations of videos shared on YouTube. It was created by Cheng, Dale and Liu at Simon Fraser University. Please see this link for details about how the data was collected.

You wish to run the following regression:

$$ln(\text{views}) = \beta_0 + \beta_1 \text{rate} + \beta_3 \text{length}$$

The variables are as follows:

- `views`: the number of views by YouTube users.
- `rate`: This is the average of the ratings that the video received. You may think of this as a proxy for video quality. (Notice that this is different from the variable `ratings` which is a count of the total number of ratings that a video has received.)
- `length:` the duration of the video in seconds.

1. Evaluate the **IID** assumption.
2. Evaluate the **No perfect Colinearity** assumption.
3. Evaluate the **Linear Conditional Expectation:** assumption.
4. Evaluate the **Homoskedastic Errors:** assumption.
5. Evaluate the **Normally Distributed Errors:** assumption.

Part 3 - Final Report

Your final report should document your analysis, communicating your findings in a way that is technically precise, clear, and persuasive.

The maximum length is 4 pages using standard pdf_document output in RStudio, and including all tables, appendices, and references. This limit is strict.

The exact format of your report is flexible (form follows function), but it should include the following elements.

1. An Introduction

Your introduction should present a research question and motivate its importance. It should draw the reader's attention to specific X and Y concepts in a way that makes the reader care about them. After reading the introduction, the reader should be prepared to understand why the models are constructed the way that they are. It is not enough to simply say, "We are looking for product features that enhance product success." Your introduction must do work for you, focusing the reader on a specific measurement goal, making them care about it, and propelling the narrative forward. This is also a good time to put your work into context, discuss cross-cutting issues, and assess the overall appropriateness of the data.

2. A Brief Description of the Data

You should assume that your reader is not familiar with the data you are using. Provide basic information such as the organization that collected the data, whether it is experimental or observational, and how units of observation were selected,.

3. A Discussion of How Key Concepts are Operationalized

You should explain which variables are used to represent your X and your Y, and how well they match these concepts. Identify key gaps between the conceptual and operational definitions. If there are alternative variables that you considered, highlight them and explain how you made your decision.

4. An Explanation of Key Modeling Decisions

1. How many observations were removed from the data, and for what reasons?
2. What transformations did you apply to your variables and why? Are they supported by scatterplots, statistical tests, or existing theory?
3. Are there covariates that were intentionally left out of your models and why? For example, did they reduce your precision too much, or are they outcome variables?

5. A Table or Visualization

You will be graded on your visual design. In particular:

1. Plots should be easy to navigate, with useful titles and axis labels.
2. Do not include raw R output. All output, including variable names, should be formatted to make it easy for an English speaker to read.
3. Plots should have a high information-to-ink ratio. If you are only communicating 2-4 numbers, a table is generally more effective than a plot.
4. Any plot or table you include must be commented on in your narrative. In other words, no output dumps!

6. A Well-Formatted Regression Table.

It is important to remember that you are not trying to create one perfect model. You will create several specifications, giving the reader a sense of how robust (or sensitive) your results are to modeling choices, and to show that you're not just cherry-picking the specification that leads to the largest effects.

You should display all of your model specifications in a regression table, using a package like `stargazer` to format your output. It should be easy for the reader to find the coefficients that represent key effects near the top of the regression table, and scan horizontally to see how they change from specification to specification. Make sure that you display the most appropriate standard errors in your table.

As you select your model specification, your goal is to encircle the space of reasonable modeling choices, and to give an overall understanding of how these choices impact results. You should strive to make your models different from each other. However, each individual model must be defensible.

At a minimum, you need to estimate at least three model specifications.

The first model you create should include *only the key variables* you want to measure. These variables might be transformed, as determined by your EDA, but the model should include the absolute minimum number of covariates (usually zero or one covariate that is so crucial it would be unreasonable to omit it).

The structure of the other models is more flexible. Most often, you will see researchers add a block of covariates from one model to the next. Each model should be defensible, and should continue to tell the story of how product features contribute to product success. This might mean including additional covariates to remove omitted variable bias; or, instead, it might mean estimating a model that operationalizes your X or Y in a different way (be sure the operationalization is substantially different). You may also create a model tailored to investigating a heterogeneous effect.

7. A Discussion of Results

In your text, comment on both *statistical significance and practical significance*. You may want to include statistical tests besides the standard t-tests for regression coefficients. Here, it is important that you make clear to your audience the practical significance of any model results. How should the product change as a result of what you have discovered? Are there limits to how much change you are proposing? What are the most important results that you have discovered, and what are the least important?

8. A Discussion of Limitations

8a. Statistical limitations of your model Make sure to evaluate all of the large sample model assumptions (or the CLM if you have a small sample). However, you do not necessarily want to discuss every assumption in your report. Instead, highlight any assumption that might pose significant problems for your analysis. For any violations that you identify, describe the statistical consequences. If you are able to identify any strategies to mitigate the consequences, explain these strategies.

Note that you may need to change your model specifications in response to violations of the large sample model.

8b. Structural limitations of your model What are the most important *omitted variables* that you were not able to include? For each variable you name, *reason about the direction of bias* caused by omitting this variable and whether the omission of this variable calls into question the core results you are reporting.

Is there a possibility of reverse causality? If so, *reason about the direction of bias* this causes.

Are there any outcome variables on the right hand side? If so, *reason about the direction of bias* this causes.

9. A Conclusion

Make sure that you end your report with a discussion that distills key insights from your work, addresses your research question, and leaves the reader with a sense of broader context or impact.

Part 3 - Final Presentation

During the Unit 14 live session, each team will give a slide presentation of their work to their classmates, who will be seated with you as collaborating data scientists. As collaborating data scientists, your classmates will need to be informed of the specific product and research question that you are addressing.

Presentation Guidelines

- **Plan for a 10 minute presentation.** Please note that this is an *incredibly* limited amount of time to present. A good rule of thumb is to use a maximum of 5 slides.
- If time is available, an additional 5 minutes will be devoted to questions.
- Begin by setting up your research question. It is quite alright to have a slide that bluntly states: "**Research Question**: Do shorter lanyards increase the accuracy of catapult launches?" (2 minutes max)
- You should ground the audience in an understanding of the data. Explain enough about your key variables that your audience can reason about your models and results. (2-3 minutes max)
- Do not present R code, discuss data wrangling, or normality - details like this are best left to the full analysis. It is tempting to want to share these process based stories with your peers, but save that time for after the presentation.
- It is a good practice to show your final regression table on a slide by itself. If you show a regression table, you need to provide your audience with enough time to digest it (minimum 2 minutes). For any table (or plot) that you show, you should minimally interpret the variables (or axes) and the key point that you are making with that piece of evidence.

Finally, a few more general thoughts:

- Practice your talk with a timer!
- If you divide your talk with your teammates, practice your section with a timer to make sure you do not talk into your teammates' time. We would hate to cut your group off before a teammate has a chance to talk.