Lab 2: Describing Generations

w203: Statistics for Data Science

Learning Objectives

- Understand and pursue the goals of a *descriptive* analysis
- Correctly evaluate the large sample regression model assumptions
- Use and interpret common variable transformations
- Evaluate both statistical and practical significance of results

Introduction

You are part of a group tasked with understanding how people of different generations differ from each other. Your task is to select a research question and a public dataset aligned with this goal, and generate a 3-page regression analysis. Your research question must specify an X concept and a Y concept. Because of the setting, the X concept must be age; but you should choose some Y concept that you are interested in. You will need to ensure that your dataset includes variables you can use to operationalize age and your Y concept of choice.

Your research question must be purely **descriptive**. Among other things, this means that your introduction must explain why understanding the relationship in question is valuable. You should not say, for example, that a company might like to understand how big each person's vacation budget is to better target promotions, since that would require a predictive model.

A common concern with descriptive questions is whether the causal version is simply more interesting (because we are ultimately interested in changing the X variable based on the analysis), or if a decision maker will misinterpret the coefficients as causal. We have purposefully chosen a setting in which there is no obvious causal model to apply. You cannot say that age *causes* some characteristic or behavior, because there is no way to manipulate an individual's age.

If your data set is large enough, you should 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. You should strive to swap the exploration set out for the confirmation set as late in the process as possible, to avoid making decisions based on the confirmation set. All numbers in your report, as well as your discussion and conclusions, should be based on your confirmation set.

Data

Your data must meet the following requirements:

- Data should be cross-sectional (i.e. each person must have one row of data, not multiple measurements taken at different times). If you find a panel dataset, you may subset a single cross section for this lab. If you have a single measurement for each person, but different people are measured at different times, that is ok, but you will typically want to include a time trend or time fixed effects to account for how time periods are different from each other (talk to your instructor if this is the case for you).

- 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 variable that you use should be metric. As you now know, it is innappropriate to use an ordinal variable as a regression outcome. However, if there is an ordinal variable that you are greatly interested in, you may ask your instructor for permission to use it. If using an ordinal variable, clearly highlight this limitation in your report.
- A binary outcome variable does not violate any large sample regression model assumptions. You are permitted to use a binary outcome, though the use of regression in this way is often frowned upon (because logistic regression is better tailored to binary outcomes).

The following sources of data are recommended:

- **General Social Survey (GSS)**. Use the Data Explorer to search for variables, or see a list in the Quick Guide
- **American Community Survey (ACS)**. Access data from the Census Website and see the list of variables.
- **Current Population Survey (CPS)**. In particular, see the Annual Social and Economic Supplement
- **Pew Research Center**. See the list of surveys on various topics

If you have a specific topic you are interested in, we encourage to find your own data that is not on the list.

Modeling

You are to create a set of regression models aligned with the top-level goal of description. A minimum of two models is required.

1. The first model should be the simplest, likely including only Age in nominal (un-transformed) form. The purpose of the model here is to provide a single number representing the average strength of the relationship. You may also use this model to test the hypothesis that there is no overall linear relationship.
2. Additional models must change how age is entered into the regression, and may include covariates or interaction terms. The main purpose is to better describe the relationship in question, leading to more human understanding or insight. You may use a transformation (e.g. a log) or a polynomial to better capture the shape of the relationship. You may also use indicator variables (e.g. age > 18) to capture discontinuities and test whether they exist. A secondary purpose of your additional models is to explore the robustness of your results. For example, does a relationship disappear when you include certain covariates? By presenting a range of models, you are helping your reader to understand if the results you found are only seen for specific modeling decisions, or if they appear across a range of reasonable models. You should strive to make your models different from each other. However, each individual model must be defensible.

Deliverable 1: A Research Proposal

Due week 11, one submission per team

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 two 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 Y?

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), and the feedback that you will receive will be brief. Your project will be given either a *green*, *yellow* or *red* light.

Projects that have been given a green-light can proceed with the question as formed, using the data as proposed. There will, quite naturally, be a lot of maturation that occurs over the course of the project, but a project that has received a green light can, and should, continue with their work.

Projects that have been given a yellow-light can proceed with caution, and changes. Something about the proposal has raised a cause for concern in the eyes of the instructor. The question that the team has proposed might be too vague for a model-based answer, or you may have proposed a classic "fishing expedition" question (i.e. "Which of these features is the most important) which is not amenable to statistical analysis, or you might have proposed a data set that is known to be problematic. This evaluation will be accompanied by feedback from the instructor about what the team should do in order to move this project onto a path for success.

A red light indicates that the team may have to start over from scratch with a new dataset. These are projects that, in the eyes of the instructor, seem likely to cause insurmountable problems for the team later on. Of course, the team may see a path to success with this type of project, but that path has not been made clear to the instructor when they read the team's proposal.

Deliverable 2: Final Report

Due week 13, one submission per team

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

The maximum length is 3 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 how people of different ages are different." Your introduction must do work for you, focusing the reader on a specific relationship, 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?

5. A Visualization

You must include one visualization that highlights the relationship between your X and your Y variable. Include a visual representation of your fitted model that showcases one of your functional forms. For example, if you take the log of age, you will plot a logarithmic curve that traces the predicted value for different values of age.

You will be graded on your overall 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, providing different views of the relationship, and also giving the reader a sense of how robust (or sensitive) your results are to modeling choices.

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. Make sure to set the significance cutoffs so that a `\*` corresponds to $p < .05$.

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. Comment on the direction and magnitude of your results, placing them in context so the reader can understand if they are important.

8. A Discussion of Limitations

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.

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.

Deliverable 3: Final Presentation

Presented week 13, one presentation per team

We will set aside time in live session for each team to present brief highlights from the work they did.

- **Plan for a 5 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 3 slides.
- Please practice your presentation with a timer, so you do not go over 5 minutes.
- 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.
- Presentations will be graded only for completeness

Deliverable 4: Within-Team Review

Due week 13, one submission per person

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 weakness, and the team's process in the project.
  - Where has 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 could the team have done 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.

The teaching team will never reveal that a specific student wrote something about their teammates without permission. However, instructors may alter a student's grade if a pattern emerges across several reviews that indicates a problem.

This reflection is due at the conclusion of your lab-work, in Gradescope, and requires one submission per person.