Lab 2: Describing a Bivariate Relationship

w203: Statistics for Data Science

Learning Objectives

- Articulate and motivate a research question aligned with description
- Discuss and justify how concepts are operationalized
- Correctly evaluate the large sample regression model assumptions
- Build a bivariate regression model by iteratively applying variable transformations
- Evaluate both statistical and practical significance of results

Introduction

Description refers to the process of representing statistical patterns in a compact, human-understandable way, in order to gain insight. Although more attention is usually given to the alternate modes of prediction and explanation, description remains important in many domains, including economics, marketing, and political science. In this lab, you will generate a short regression analysis.

Your first task is to select a descriptive research question and a public dataset that you can use to address it. To constrain the scale of the project, you must select a single X concept and a single Y concept. You will need to ensure that your dataset includes variables you can use to operationalize both concepts.

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. A common error is to motivate the research question in terms of prediction or explanation. Here are two examples of this type of error:

- One writes that a company might like to understand how big each person's vacation budget is to better target promotions. Since the objective requires an accurate prediction of Y for each value of X, this requires a predictive model.
- One writes that medical professionals would like to understand how to decrease heart attack risk. Since this objective involves manipulating something in the real world, this requires an explanatory model.

In a descriptive mode, we seek to understand, quite simply, what a distribution *is*. You must therefore present reasons that someone would want to understand what the relationship between your X and your Y is. Some classic examples of descriptive questions follow:

- What is the wage gap between people of different genders? Understanding what the gap is is valuable as a measure of fairness in society.
- How much do people with different political leanings care about the economy? This information would be useful to design political campaigns.
- How much do consumers prefer local brands over other brands? This question was studied by Bronnenberg et al., [1] extracting insights that are useful to marketing teams.

Since you are performing a bivariate analysis, you may think of description as revealing the "shape" of the relationship between your X and your Y variable and making it human understandable.

Data

You must find you own public data source for this lab. Your data must meet the following requirements:

- Data should be cross-sectional; that is, each person (or other unit) 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.
- Both your X and Y variable should be metric with more than two levels. 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.

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. Ideally, all modeling decisions - including rules for what data to code NA, how to transform variables, and what tests to run - are made using only the exploration set. Only after your report is nearly done, should you swap to the confirmation set and recalculate all the numbers in your report, including summaries, model coefficients, and p-values. Your discussion and conclusions should be based on these final numbers from your confirmation set.

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 two regression models aligned with the top-level goal of description.

1. The first model should be the simplest, with both variables untransformed (in nominal 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. The second model must change how your X and/or your Y variable is entered into the regression. The main purpose is to more closely describe the shape of 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.

When creating multiple models, you should strive to make them different from each other. However, each individual model must be defensible.

Final Report

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

Page limits:

- Main report: 3 pages
- Appendix: 1 additional page

You must meet these page limits using standard pdf_document output in RStudio. They include all tables, appendices, and references. These limits are strict.

The one-page appendix is intended for extra information that will help your instructor assess your model building process. Please include the following elements:

1. A Link to your Data Source If you used specialized code to access your data, please include that here. Please make sure your instructor has the ability to access the data.

2. A List of Model Specifications you Tried We are interested in seeing how you arrived at your final model. In just a sentence, please provide a reason or something that you learned from each specification.

3. A Residuals-vs-Fitted-values Plot Please generate this plot for your final model that includes variable transformations. Your instructor will use this plot to assess how well you have captured the shape of the relationship between your X and your Y variable. For example, if there is a clear parabolic pattern in your residuals-vs-fitted plot, that is a signal that you should have included a square term.

Evaluation Criteria

We present the following criteria to guide you to a professional-quality report. Moreover, these criteria are also the ones we will use to grade your report. The descriptions below are copied directly from our grading rubric.

1. Introduction

An introduction that is scored in the top level has very successfully made the case for the study. It will have explained why the topic is interesting, and provided compelling reasons to care about not just the general area, but rather about every concept in the research question and the statistical results to be generated. The introduction will be engaging from the very first sentence, and create a logical story that leads the reader step-by-step to the research question. After reading the introduction, no part of the research question will appear arbitrary or unexpected, instead flowing naturally from the background provided.

2. Description of the Data Source

A report that is scored in the top level will describe the provenance of the data; the audience will know the source of the data, the method used to collect the data, the units of observation of the data, and important features of the data that are useful for judging the analysis.

3. Data Wrangling

A report that is scored in the top level on data wrangling will have succeeded to produce a modern, legible data pipeline from raw data to data for analysis. Because there are many pieces of data that are being marshaled, the reports that score in the top level will have refactored different steps in the data handling into separate files, functions, or other units. It should be clear what, and how, any additional features are derived from this data. The analysis avoid defining additional data frames when a single source of truth data would suffice.

4. Operationalization

A report that is scored in the top level on operationalization will have precisely articulated and justified the decisions leading to the variables used in the analysis. The reader will be left with a clear understanding of the concepts in the research question, and how they relate to the operational definitions, including any major gaps that may impact the interpretation of results. You should also list how many observations you remove and for what reasons. When there is more than one reasonable way to operationalize a concept, the report will explain the alternatives and provide reasons for the one that was selected. Note that there is often more than one way to operationalize a concept; we are less interested in whether you make the best possible choice, and more in how well you defend your decisions.

5. A Visualization

You are required to include at least one plot or table in your main report that highlights the relationship between your X and your Y variable (plots in the appendix do not count). Include a visual representation of your final model predictions on the plot. A report that is scored in the top level will have plots that effectively transmit information, engage the reader's interest, maximize usability, and follow best practices of data visualization. Titles and labels will be informative and written in plain english, avoiding variable names or other artifacts of R code. Plots will have a good ratio of information to space or information to ink; a large or complicated plot will not be used when simple plot or table would show the same information more directly. Axis limits will be chosen to minimize visual distortion and avoid misleading the viewer. Plots will be free of visual artifacts created by binning. Colors and line types will be chosen to reinforce the meanings of variable levels and with thought given to accessibility for the visually-impaired. Your report will be free of "output dumps" - code output that has not been formatted for human-readability. Every single plot and table in the report will be discussed in the narrative of the report.

6. Model Specification

A report that is scored in the top level will have chosen a set of regression models that strongly support the goal of the study. Variables transformations will be chosen to inform the reader of the shape of the joint distribution, and will be human-understandable. A reason will be provided for the chosen variable transformations. Ordinal variables will not be treated as metric. All model specifications will be displayed in a regression table, using a package like `stargazer` to format your output. Displayed standard errors will be correctly chosen. Significance cutoffs will be set so that a `\*` corresponds to $p < .05$.

7. Model Assumptions

A report that scores in the top-level has provided an thorough and precise assessment of the assumptions supporting the regression. The list of assumptions is appropriate given the sample size and modeling goals. Each assumption is evaluated fairly, and discussed defensively - without overstating how credible the assumption is or rendering a final up-or-down judgement on the assumption. The report will not miss any important violations of any assumption. Where possible, the report discusses the consequences for the analysis of a violated assumption.

8. Model Results and Interpretation

A report that scores in the top level will correctly interpret statistical significance, clearly interpret practical significance, and comment on the broader implications of the results. It may want to include statistical tests besides the standard t-tests for regression coefficients. When discussing practical significance, comment on both the direction and magnitude of your coefficients, placing them in context so the reader can understand if they are important. To help the reader understand your fitted model, you may want to describe hypothetical datapoints (e.g. a hypothetical person with 1 cat is predicted to spend $2400 on pet care. That rises to $3200 for a hypothetical person with 2 cats...).

9. Overall Effect

A report that scores in the top level will have met expectations for professionalism in data-based writing, reasoning and argument for this point in the course. It can be presented, as is, to another student in the course, and that student could read, interpret and take away the aims, intents, and conclusions of the report.

[1] Bronnenberg, B. J., S. K. Dhar, J.-P. H. Dubé. 2009. Brand history, geography, and the persistence of CPG brand shares. J. Political Econom. 117(1) 87-115