Lab 1: Linear Regression in R

Joshua Klugman

1/13/2021

In this assignment, you will fit a linear regression model in R. You should use your own dataset. You either should have a continuous outcome or you should be comfortable with running a Linear Probability Model (LPM) and you will have at least two predictors; ideally at least one at “level 1” and at least one at “level 2”.

You will include one appendix containing an R console file that will list your data management steps and your analyses. This file should be run from a cleaned and organized R script that does everything that shows up in the assignment.

When grading your assignment, I will focus on your professional presentation, your understanding of regression, your ability to convey that understanding to your readers who may not have taken this class, and your ability to follow directions.

To assist in my grading, please answer these questions in order with section headers letting me know which element you are addressing.

|  |  |
| --- | --- |
| Points | Element |
| 5 | **Research question.** You will briefly describe what question you are trying to answer with this analysis. |
| 10 | **Data source.** You will describe your data source and make clear **where**, **when** and **how** your cases were selected. To what extent were they randomly sampled? |
|  | How was data collected from individuals (e.g. online survey, phone interview, in-person diagnostic assessment)? |
|  | If you are using an experiment, what were the procedures used for the treatment/control conditions, and how were subjects assigned to those different conditions? |
|  | If you are using longitudinal data you will be clear about the timing of follow-ups. |
|  | Do not get around answering these questions by citing some other article. |
| 5 | **Dropped Cases.** How many cases do you have before you drop cases due to scope conditions? |
|  | How many cases do you have after you drop cases due to scope conditions? |
|  | How many cases do you have after you drop cases due to missingness (i.e. subject noncompliance)? |
| 10 | **Variables.** You will describe each of your variables. |
|  | Make clear the role they play in your analysis (outcome, focal predictor, control variable, mediator, moderator, etc.). |
|  | Briefly describe how they were measured. |
|  | Their valence and unit of measurement should be clear. |
|  | Are you using any standardizations or transformations? |
| 5 | **Scatterplot matrix.** Present a scatterplot matrix of your outcome and your predictors using the ggpairs function. |
|  | Suggested order of variables: continuous predictors, outcome, then categorical predictors. |
|  | Categorical variables should be treated as factor variables |
|  | You may want to opt for cryptic one-letter category labels so the graph is legible. |
| 5 | **Distribution of Outcome.** You present a histogram of your outcome using ggplot2. |
|  | If the outcome is binary or severely skewed articulate the problems with analyzing a skewed outcome. |
| 10 | **Summary Statistics.** Create a professional table of summary statistics (not R output), giving the mean and standard deviation of variables. |
|  | ALL variables should be presented the way they are entered into your regression models–if you standardize/transform/center a variable, you should present summary statistics on the standardized/transformed variable (it is optional to present summary statistics on the raw values). |
|  | No cryptic variable names. |
|  | Variable names should clearly indicate the valence of the variable, and no dummy variable names like “gender” or “race”. |
|  | Acronyms should be explained in a note. |
|  | Do not just copy and paste tables directly from a statistical package in your paper. |
|  | The unit of measurement & valence should be clear, especially if the variable lacks a concrete metric or if it is z-standardized (with a “z” or “SD”). |
|  | Everything should be clearly labeled and titled. |
| 10 | **Regression table.** Prepare a professional table showing the associations between the outcome, on the one hand, and at least your **key** predictors. |
|  | The model(s) you present should speak to your substantive concerns. |
|  | Make sure you indicate z-standardization / transformations (including for the outcome). |
|  | Make sure the valence of all variables are clear. |
|  | If you decide to omit control variables from your table, you should add a note listing them. |
|  | I should be able to read your table by itself and interpret all main effects and interaction effects (if any) without consulting your text. |
|  | Your presentation should be compact–you should give me the slope and standard errors and stars (do not present the standard error AND the *t*-statistic!). |
| 20 | **Write-Up**. Write up your results, as you would in a journal article. If you have a lot of predictors, choose a couple of predictors to focus on. Your write-up should give a sense of the statistical & substantive significance of your effects. |
| 10 | **Graph**. Graph the effects of a variable using either a line graph or a bar graph using ggplot2. |
|  | The graph should be professional looking, with a title, axis labels, a y-axis that does not exaggerate the effect of the predictor, and no cryptic variable names. |
|  | There should be no ambiguity about units of measurement or valence. |
|  | You should present error bars or ribbons to indicate the precision of the estimate. |
| 10 | **Apendix**. Your R syntax that is organized and clearly labeled. There is a logical progression from reading in data, cleaning it, creating variables, and running analyses. |
| 0 - -100 | **Glaring Errors**. There will be no glaring errors that indicate sloppiness or send a signal that you do not know what you are doing (e.g. misinterpreting coefficients). There will be consistency among all elements of the assignment. |

**Upload your assignment to Canvas. Check your submission to make sure there are no issues with formatting. If there are email the assignment to me (**[**klugman@temple.edu**](mailto:klugman@temple.edu)**).**