Lab 2: Regression to Study the Spread of Covid-19

w203: Statistics for Data Science

10/28/2020

Introduction

In this lab, you will apply what you are learning about linear regression to study the spread of COVID-19. Your task is to select a research question, then conduct a regression study to analyze it.

Your research question must focus attention on a specific measurement goal. It is not enough to say "we are looking for policies that help stop COVID-19" (That would be a fishing expedition). Instead, use your introduction to motivate a specific effect for measurement.

Once you have a measurement goal, you will build a set of linear models that are tailored to that goal, and display them in a well formatted regression table. You will finally use your conclusion to distill important conclusions from your estimates.

This is a group assignment. Your live session instructor will coordinate the formation of groups. We would like to encourage teams to focus on using the lab as a way 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. As a result, we encourage teams to allow individuals to take risks and be supportive in the face of successes and failures. Create an opportunity for people who want to improve a particular skill to do so – this might be project coordination, management of code through git, plotting, or any of the many aspects that you'll work on. *We hope that you can support and learn from one another through this team-based project.*

Deliverables

Deliverable Name	Week Due	Grade Weight
Draft Report	Week 12	10%
Within-Team Review	Week 12	5%
Cross-Team Review	Week 13	10%
Final Report	Week 14	65%
Final Presentation	Week 14	10%

The Data

The data is provided in a file within this repository. Majid Maki-Nayeri compiled the data, drawing many variables from the COVID-19 US state policy database (Raifman J, Nocka K, Jones D, Bor J, Lipson S, Jay J, and Chan P.). Supporting documents like the specific legal language used by states, additional data sources, and much more are available in unstructured format here.

The dataset includes:

1. Variables representing the spread of the disease;
2. Variables representing state-level policy responses; and,
3. General state-level characteristics.

If you want to, you are allowed to add extra variables from external sources. However, this is not necessary, and we will not assign any bonus points to teams that derive unique data, as our focus is on statistics and statistical writing.

Final Project Components

Draft Report

In the first stage of the lab, you will create a draft report. You should aim to make this report as complete as possible, so that you get the best possible feedback from the Cross-Team Review.

The exact format of your report is flexible, but it should include the following elements.

1. An Introduction

Your introduction should present a research question and explain the concept that you're attempting to measure and how it will be operationalized. This section should pave the way for the body of the report, preparing the reader to understand why the models are constructed the way that they are. It is not enough to simply say "We are looking for policies that help against COVID" 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 good time to put your work into context, discuss cross-cutting issues, and assess the overall appropriateness of the data.

2. A Model Building Process

You will next build a set of models to investigate your research question, documenting your decisions. Here are some things to keep in mind during your model building process:

1. *What do you want to measure*? Make sure you identify one, or a few, variables that will allow you to derive conclusions relevant to your research question, and include those variables in all model specifications.
2. Is your modeling goal one of description or explanation?
3. What covariates help you achieve your modeling goals? What covariates are problematic, either due to *collinearity*, or because they are outcomes that will absorb some of a causal effect you want to measure?
4. What *transformations*, if any, should you apply to each variable? These transformations might reveal linearities in scatterplots, make your results relevant, or help you meet model assumptions.
5. Are your choices supported by exploratory data analysis (*EDA*)? You will likely start with some general EDA to *detect anomalies* (missing values, top-coded variables, etc.). From then on, your EDA should be interspersed with your model building. Use visual tools to *guide* your decisions. You can also leverage statistical *tests* to help assess whether variables, or groups of variables, are improving model fit.

At the same time, 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.

At a minimum, you should include the following three specifications:

1. **Model 1**: One model with *only the key variables* you want to measure (possibly transformed, as determined by your EDA), and no other covariates (or perhaps one, or at most two, covariates if they are so crucial that it would be unreasonable to omit them)
2. **Model 2**: One model that includes *key explanatory variables and covariates that you believe advance your modeling* goals without introducing too much collinearity or causing other issues. This model should strike a balance between accuracy and parsimony and reflect your best understanding of the relationships among key variables.
3. **Model 3**: One model that includes the *previous covariates, and many other covariates*, erring on the side of inclusion. A key purpose of this model is to demonstrate the robustness of your results to model specification. (However, you should still not include variables that are clearly unreasonable. For example, don't include outcome variables that will absorb some of the causal effect you are interested in measuring.)

Guided by your background knowledge and your EDA, other specifications may make sense. You are trying to choose points that encircle the space of reasonable modeling choices, to give an overall understanding of how these choices impact results.

3. Limitations of your Model

As a team, evaluate all of the CLM assumptions that must hold for your models. However, do not report an exhaustive examination all 5 CLM assumption. Instead, bring forward only those assumptions that you think pose significant problems for your analysis. For each problem 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 CLM.

4. A Regression Table

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, along with significance stars.

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.

5. Discussion of Omitted Variables

If your team has chosen an explanatory (i.e. causal) question to evaluate, then identify what you think are the 5 most important *omitted variables* that bias the coefficients you care about. For each variable, you should *reason about the direction of bias* caused by omitting this variable. If you can argue whether the bias is large or small, that is even better. State whether you have any variables available that may proxy (even imperfectly) for the omitted variable. Pay particular attention to whether each omitted variable bias is *towards zero or away from zero*. You will use this information to judge whether the effects you find are likely to be real, or whether they might be entirely an artifact of omitted variable bias.

6. Conclusion

Make sure that you end your report with a discussion that distills key-takeaways from your estimates, addresses your research question, and draws attention to larger contexts.

Submission

- Submit your draft report via ISVC; please do not submit via email.
- Submit 2 files:
    1. A pdf, html or md file including the summary, the details of your analysis, and all the R codes used to produce the analysis. *Please show code in your compiled file.*
    2. The Rmd or source file used to produce the pdf file.
- Only one group member needs to submit the files.
- Be sure to include the names of all team members in your report. Place the word 'draft' in the file names.
- Please limit your submission to 8000 words, excluding code cells and R output.

Within-Team Review

Being an effective team member is a crucial part of data science work. Your performance on 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 to 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. 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 me know in this evaluation. As well, if there are any individual performances that require special attention, please also let me know. I will treat these reviews as confidential and will not take any action without first consulting you.

You will submit this through ISVC, and like all parts of your educational record, this will be treated confidentially by the instructional team.

Cross-Team Review

In the cross-team review, you will provide feedback on another team's draft report. We will ask you to comment separately on different sections. The following list is very similar to the rubric we will use when grading your final report.

- **Introduction.** Is the introduction clear? Is the research question specific and well defined? Does the introduction motivate a specific concept to be measured and explain how it will be operationalized. Does it do a good job of preparing the reader to understand the model specifications?
- **The Initial Data Loading and Cleaning.** Did the team notice any anomalous values? Is there a sufficient justification for any data points that are removed? Did the report note any coding features that affect the meaning of variables (e.g. top-coding or bottom-coding)? Overall, does the report demonstrate a thorough understanding of the data?
- **The Model Building Process.** Overall, is each step in the model building process supported by EDA? Is the outcome variable (or variables) appropriate? Is there a thorough univariate analysis of the outcome variable? Did the team identify one, or very small number of explanatory variables and perform a thorough univariate analysis of each one? Did the team clearly state why they chose these

explanatory variables, does this explanation make sense in term of their research question? Did the team consider available variable transformations and select them with an eye towards model plausibility and interpretability? Are transformations used to expose linear relationships in scatterplots? Is there enough explanation in the text to understand the meaning of each visualization?

- **Regression Models:**
  - **Model 1** Does this model only include key explanatory variables? Do the variables make sense given the measurement goals? Did the team apply reasonable transformations to these variables, to capture the nature of the relationships?
  - **Model 2** Does this model represent a balanced approach, including variables that advance modeling goals without causing major issues? Does the model succeed in reducing standard errors of the key variables compared to the base model? Does it capture major non-linearities in the joint distribution of the variables?
  - **Model 3** Does this model represent a maximalist approach, erring on the side of including most variables? Is it still a reasonable model? Are there any variables that are outcomes, and should therefore still be excluded? Is there too much multicolinearity, to the point that the key causal effects cannot be measured?
- **Plots, Figures, and Table** Do the plots, figures and tables that the team has chosen to include successfully move forward the argument that they are making? Do they have a good ratio of "Information to Ink" (Tufte)? Has the team chosen the most effective method (a table or a chart) to display their evidence? Is that table or chart as communicative as it can be? Is every single plot, figure, or table that is included in the report referenced in the main text?
- **Assessment of the CLM.** Has the team assessed each of the CLM assumptions (including random sampling)? Did they use visual tools or statistical tests, as appropriate? Did they respond appropriately to any violations?
- **A Regression Table.** Are the model specifications properly chosen to outline the boundary of reasonable choices? Is it easy to find key coefficients in the regression table? Does the text include a discussion of practical significance for key effects?
- **An Omitted Variables Discussion.** Did the report miss any important sources of omitted variable bias? Are the estimated directions of bias correct? Was their explanation clear? Is the discussion connected to whether the key effects are real or whether they may be solely an artifact of omitted variable bias?
- **Conclusion.** Does the conclusion address the research question? Does it raise interesting points beyond numerical estimates? Does it place relevant context around the results?
- Are there any other errors, faulty logic, unclear or unpersuasive writing, or other elements that leave you less convinced by the conclusions?

Please be thorough and read the report critically, actively trying to find areas that the report could be improved. Your comments will directly help your peers get the most value out of the project.

Final Report

In the final stage of the project, you will incorporate the feedback you receive, and use what you've learned about OLS inference to create a final report.

We will assess your final report using a rubric that includes the elements listed above. We will also consider whether you have correctly applied the statistical techniques in your report.

Please limit your submission to 8000 words, excluding code cells and R output.

As above, you must submit both the source and pdf files. Be sure to include the names of all team members in your report. Place the word 'final' in the file names.

Final Presentation

During the Unit 14 live session, each team will give a slide presentation of their work to their classmates – i.e. collaborating data scientists. This audience is generally aware of the project that you're working on, but they will need to be reminded of the specific research question that you are addressing. **The time limit for your slide presentation is 10 minutes.** We will budget a few extra minutes for questions after you are done. We'd like to emphasize that this is an *incredibly* limited amount of time to present. The materials that you present should reflect these serious constraints!

1. There should be no more than two slides – probably one is best – that set-up your research problem and these slides should take no more than two minutes to present. (1 minute)
2. There should be at least one, and not more than two, slides that describe the one or two most important variables that you're using in your models. These slides should cover the important features of the variables that you're using: how are they are measured, the unit of observation, and why these *particular* variables are appropriate to use to answer your research question. (3 minutes)
    - Do not present R code, discuss data wrangling, or discuss normality - details like this are best left to the full analysis.
3. There should then be several slides that provide what you've learned from your models. If you show model results, you need to provide your audience with enough time to read and engage with these models; not flash past them. Any model you show will take at least one minute to talk about.

Finally, a few more general thoughts:

- Practice your talk with a timer. We saw the 2020 debates, and we're going to end your talk 10 minutes in – we have a mute button! :joy:
- If you divide your talk with your teammates, practice your section with a timer so that you do not spill over into your teammates' time.
- We *strongly* recommend having no more than 5 slides total.