Lab 3: A Regression Study of COVID-19

Peer Feedback for the regression model created by:
Adam Sayre, Brodric Cormier, Sanjay Elangovan

Feedback provided by:
Eric Lundy, Gabriel Ohaike, Javed Roshan

Overall Feedback:

1.0 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?

Feedback] Introduction is comprehensive. The research question is not explicitly defined. We understand the investigation being undertaken. We recommend an explicit question definition.

We also recommend operationalizing the core variables planned to be used in the model by further defining their relationship to the research question.

2.0 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?

Feedback]

The team did a callout on the following data anomalies:

- "ReopenBusiness" have data populated and "Closed_NE_Business_data" is empty
- Two rows for Arizona state

We, however, recommend keeping at least one of the "Arizona" data rows. Couple of options to consider:

- Validate the data by further researching external sources
- Merge the rows using mean as applicable

3.0 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 a few, 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 interoperability? Are transformations used to expose linear relationships in scatterplots? Is there enough explanation in the text to understand the meaning of each visualization?

Feedback]

We recommend rationalizing the use of additional variables in the improvement v1 and v2 models.

If you have considered transforming variables in the v1 and v2 models, please do state them in the report.

4.0 Regression Models: Base Model. 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?

Feedback]

The team did a good job in incrementally building the versions of the model.

Taking the feedback of better defining the research question will allow for a better connection between variables and their measurement goals.

Looking at the scatter plots, a log of Total Cases might improve the model. Please consider it.

4.1 Regression Model: Second Model. 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 nonlinearities in the joint distribution of the variables?

Feedback]

Please include code that defines how the heteroskedasticity is addressed by using robust standard errors.

4.2 Regression Model: Third Model. 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 multicollinearity, to the point that the key causal effects cannot be measured?

Feedback]

Yes. The 3rd model addresses above questions.

4.3 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?

Feedback]

Please state clearly if Zero Conditional Mean assumption is violated or not.

4.4 The 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?

Feedback]

Please consider showing inclusion of the standard errors in the Regression Table. Example below:

se_model <- sqrt(diag(vcovHC(model))) #Needed for heteroskedasticity

5.0 The 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?

Feedback]

Please consider defining omitted variable bias in the model (eg:- $$\beta_4 * \alpha_1$$) as well.

6.0 Conclusion. Does the conclusion address the research question? Does it raise interesting points beyond numerical estimates? Does it place relevant context around the results?

Feedback]

Please consider adding a statement about practical significance in this section.

7.0 Can you find any other errors, faulty logic, unclear or unpersuasive writing, or other elements that leave you less convinced by the conclusions?

Feedback]

The practical significance considered R^2. However, please consider using adjusted R^2 measure instead. Looking at the stargazer table, the difference in the improvement model 1 & 2 is 1% only.