Peer Review for Lab 3

Lab by E. Hulburd, S. Vadakkumkoor, and F. Vergara

Reviewed by L. Evans, D. Rasband, and Y. Zamriy

April 9, 2018

Stage 2: Peer Feedback

Instructors prompt: In Stage 2, 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

Instructors prompt: As you understand it, what is the motivation for this team's report? Does the introduction as written make the motivation easy to understand? Is the analysis well-motivated?

Peer review: The motivation for this team's report is to examine the data to come up with conclusions about which variables in which contexts have had an effect on crime rate so that an informed political candidate might try to effect changes in those areas. A few comments on the content:

- We need to be clear on who the audience of this report will be. Currently, it is written as if the intended audience is Micah; the intended audience should be the *candidate*.
- Additionally, it is stated in the introduction that the intended audience (the candidate) should interpret the report and develop initial thoughts on the policies rather than the team provide reccomendations on how to lower crime.
- Definition of terms is not clear. What does "contextualizing models" mean? "Elastic variables" usually applies to variables in the log-log model and it is unclear if this is what this term referring to
- Some spellings and grammar need attention.

The Initial EDA

Instructors prompt: Is the EDA presented in a systematic and transparent way? Did the team notice any anomalous values? Is there a sufficient justification for any datapoints that are removed? Did the report note any coding features that affect the meaning of variables (e.g. top-coding or bottom-coding)? Can you identify anything the team could do to improve its understanding or treatment of the data?

Overall comments:

- Results should be allowed to print out, rather than just shown as comments inline. This ensures accuracy of what is reported and allows for proper comparison with what is written in the prose.
- The biggest challenge with this report is lack of outputs in relevant places and supporting narrative.
- Furthermore, there are pieces of code with big outputs, also with limited commentary on the process, logic and insights.

Data Cleaning

- Could consider using na.omit to remove rows that are missing data, rather than `data <- data[-seq(92,97)]` as more succinct and reproducible style.
- Nice work on finding the duplicates.

- Once the working dataset is finalized, it would be good to present the updated dimensions, i.e. how many rows were removed.

Identifying problematic values

- Well thought-out arguments to keep the `prbarr` and `prbconv` over 1.
- Regarding `avgsen`, it represents average sentence across various crimes. Isn't it possible that there is a much greater number of short sentences (a few days) and very few long sentences (years) that average to 20 days? Note that sentencing will be handed out from county courts, rather than State or Federal.

Cross Section Variables

- The zone that is not West or Central is missed in the analysis of these dummy variables.

Exploring the Dependent Variable

- It is never actually specified that population `crmrte` denominator is the same as in the `density` numerator.
- Some more thought needs to be given to the concept of comparing an average county crime per sq. mile for the state. This logic could be strongly challenged.
- Is comparison to Neighborhood Scout valid? Are you comparing to the same state? Is it county level? A little bit more context would be helpful (along with reference to the source).
- Consideration could be given to using the rate per 1000 people in Neighborhood Scout to compare crime rate with, which does then appear reasonable.
- Regarding log transformation, the values will be negative but, hypothetically, positive values are possible as well. So this should not be a concern.

Exploring Potential independent Variable Transformations

- There is a big piece of code here without any output or comments which should be explained. Relevant charts should probably be included to support at least two transformations (`log.polpc` and `inv.prbarr`). Otherwise, there is no clear explanation as to why they were picked.
- Also why `inv.prbarr` and not `log.prbarr`? The chosen name is misleading as 'inv' suggests inverse.

Other independent variables

- There is a lot of code with the output all buried in the appendix. One paragraph summary of what was achieved would be very helpful for the reader.

Identifying possible outliers

- Now that you've identified outliers, some thought as to the origin of these outliers should be provided to decide on what action to take with them.

Correlation Analysis

- Great chart! It is a good representation of correlation matrix. Needs a title.
- Again, a summary with more than one sentence would be helpful. Otherwise, the code and the outputs take all the space without a clear purpose.

Observations from EDA

- This is a good summary, but it is mostly related to correlations, not the entire EDA. For example, why are there no comments on outliers that were identified in one of the sections?
- Consider ordering the observations from EDA interest or relevance. It might be good to pick out those that are most interesting and complementary to the models.
- The discussion around correlation noted between `taxpc` and crime rate doesn't seem to fit.

The Model Building Process

Instructors prompt: Overall, is each step in the model building process supported by EDA? Is the outcome variable (or variables) appropriate? 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?

Overall comments:

- In general, it appears that the concepts learned in the class were not applied (at least explicitly). For example, there is no hypothesis stated about causes of crime and available variables that might capture this relationship.
- The process of model building is automated based on a set of rules that have not been stated to allow for "best fit". The researchers' thoughts and logic are missing in this report due to limited high level commentary and opaque variable selection process.
- While the methods demonstrated are very practical and helpful, and we learned a lot from their application, be sure their use is within the spirit of this exercise.
- Remember to include titles and labels on charts.

Selecting Models

- Can you please include comments about the helper functions? For example, how are they relevant to the analysis. It is clear that you will use them, but their purpose is never outlined.

Stepwise Regression

- Why did you decide to do stepwise regression? What are your goals in this analysis? Are you trying to improve the prediction of crime rate? Or are you trying to explain it?
- The process is not transparent. For example, the choice of VIF of 3 or the use of only continuous variables was never explained. What are the variables that were tested? They are captured through the helper function, but how do we know that it is doing what you intended? A reader, such as a political candidate, may struggle to follow here.

Using ANOVA to Further Reduce Model Complexity

- Again, why this process? What is the goal here? What hypothesis are you testing?

Using Leaps Algorithm to Maximize Adjusted R-Squared

- This is an interesting process to figure out what variables have the best predictive power for the model. But how does it help in achieving the main objective in this analysis?
- The commentary on `prbarr` should be adjusted as it actually relates to `inv.prbarr`.

Model using log of crime rate

- Regarding arrests per capita variable. You are multiplying arrests/offenses by crimes/pop. Be confident that 'offenses' is the same as 'crimes'.
- How does this variable (arrests per capita variable) help the practical need to identify ways to reduce crimes?
- Regarding convictions per capita variable. You are multiplying convictions/arrests by crimes/pop. Confirm this really represents convictions per capita?
- Are you sure about your interpretation of the `log_crmrte`? For example, the direction of the relationship is the same: `log(0.05) < log(0.09)`.
- An interpretation of the coefficients would be useful in this section.

The Regression Table

Instructors prompt: 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?

Overall comments:

- The stargazer output has four models instead of 3. Why did you decide to keep 4 models? What are the differences between them?
- The dependent names are not stated. It is important to do so because #4 is `log_crmrte`.
- The coefficients for #4 are not comparable to the other three.
- There is no coefficient interpretation.
- There is no comparison of coefficients across models (where applicable).
- There is no discussion of practical significance of key effects.

Evaluating most influential data points

- This section has no analysis, mostly outputs that are not explained. The charts have different scale of the y axis, that could be aligned to demonstrate the difference more clearly.

Conclusion

Instructors prompt: Does the conclusion address the big-picture concerns that would be at the center of a political campaign? Does it raise interesting points beyond numerical estimates? Does it place relevant context around the results?

Overall comments:

- Conclusions read a little like results. Throw some interpretation of the models in there to back up the claims, and that's a good results section.
- Where does the conclusion that lower probability of arrest drives crime rates come from? Probability of conviction also appears favorable.
- Comprehension of the meaning of the inverse `prbarr` needs to be provided as the story has not naturally led there.

The Omitted Variables Discussion

Instructors prompt: Did the report miss any important sources of omitted variable bias? For each omitted variable, is there a complete discussion of the direction of bias? Are the estimated directions of bias correct? Does the team consider possible proxy variables, and if so do you find these choices plausible? Is the discussion

of omitted variables linked back to the presentation of main results? In other words, does the team adequately re-evaluate their estimated effects in light of the sources of bias?

Overall comments:

- The report includes a comprehensive list of potentially omitted variables.
- However, next step is to sort through them to identify those that have impacts on the outcomes of the model, and include some discussion of why they would have an impact.
- This section is only loosely related to your main findings.
- Conclusions can then take all this into account and provide the candidate with some clear recommendations for policies. It would be helpful to make these recommendations practical: more detectives, better equipment, improved relationships with local communities, etc.

Appendices

- Some of these outputs should be in the main sections.