Feedback

Evans, Rasband, Zamriy

April 8, 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:

- What is the intended audience? Currently it is written as if the intended audience is Micah, but I believe the intended audience is the *candidate*
- Additionally, intended audience (the candidate) is left to its own device to interpret the report and develop initial thoughts on the policies
- Definition of terms is not clear. What does it mean "contextualizing models"? "Elastic variables" usually applies to variables in the log-log model. Is that 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
- On the other hand, there are pieces of code with big outputs, but no relevant commentary on the process, logic and insights
- Several sentences started with "ultimately"

Data Cleaning

- Can use na.omit to remove rows that are missing data, rather than `data <- data[-seq(92,97)]`
- There's no reason really to remove `county` and `year` from the data. It can be left in and just worked around
- Nice work on finding the dupes
- What are the dimensions of the final dataset? How many rows were removed?

Identifying problematic values

- Good arguments to keep the pbarr and pconv over 1
- Regarding `avgsen`. It represents average sentence across various crimes. Isn't it possible that there is a much larger number of short sentences (a few days) and a few very long sentences (years) that average to 20 days?

Cross Section Variables

- The zone that is not West or Central is missed

Exploring the Dependent Variable

- It is never actually specified that population crmrte denominator is the same as in density numerator
- 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)
- Regarding log transformation. Yes, the values will be negative, but hypothetically positive values are possible as well. So this should not be actually a concern

Exploring Potential independent Variable Transformations

- There is a big piece of code here without any output or comments. What is it achieving? You should probably include the charts relevant to at least two transformations that you decided to include (log.polpc and inv.prbarr). Otherwise, there is no clear explanation why you picked them
- Also why `inv.prbarr` and not `log.prbarr`? The chosen name is confusing.

Other independent variables

- There is a lot of code with the output. But what did you achieve with this output? One paragraph summary would be very helpful
- Now that you've identified outliers, what are you going to do with them?

Correlation Analysis

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

Observations from EDA

- Could the observations from EDA be ordered by interest or relivance? #4 on police per capita interaction is more interesting than urban being correlated with density. 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
- This is a good summary, but it is mostly related to correlations, not the entire EDA. For example, why no comments on outliers that you identified in one of the sections?

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 interperability? 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 aribitrary rules to allow for "best fit". The researchers' thoughts and logic are missing in this report due to sparse high level commentary and obscure variable selection process
- However, we do need to state that we learned a lot from this team's work. The methods they used are actually very practical and helpful. But this report might not be the best place to apply them as we are more interested in causality and hypothesis than predictive power of the model
- Charts could do with titles and labels

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. Why VIF of 3? Why only continuous variables? 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?

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

- It's an interesting process to figure out what variables have the best predictive power for the model. But how does it help in acheiving 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/offences by crimes/pop. Are you confident that offences 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. How is it convictions per capita?

- Are you sure about your interpretation of the log_crmrte? For example, the direction of the relationship is the same: $\log(0.05) < \log(0.09)$
- There is no coefficient interpretation
- The motivation from the write-up in this section is unclear

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 influence datapoints

- This section has no analysis, mostly outputs.

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:

- Where does the conclusions 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
- Conclusions read a little like results. Throw some interpretation of the models in there to back up the claims, and that a good results section.

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:

- Comprehensive list of potentially omitted variables
- However, next step is to sort through them to identify those that have impacts to 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 an provide the candidate with some clear reccomendations 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.