W203 Lab 3 Feedback

Avinash Chandrasekaran, Saurav Datta, Deepak Nagaraj

4/8/2018

Feedback

Introduction

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?

The motivation for this report is twofold:

- Understand the relationship between crime rate and a number of independent variables
- Provide policy recommendations to local government based on analysis

We think the introduction is well-written.

It will be good to call out that the analysis will be done using ordinary least square regressions.

We have the following comments on the clean-up section:

na.omit() is dangerous to use as-is. It is better to show proof for (1) 6 rows of NA's all over, and (2) no NA's anywhere else, before using it.

More cleanup is possible: there is one duplicate record.

The EDA involves 3 pieces: a) Omitting NA's in crime b) Changing county, year, west, central, urban to factor c) Casting prbconv to numeric

Beyond the 3 EDA, the team goes on to normalize the variables. If there are outliers and extreme values, the normalization be ineffective. There could have been more data analysis to find outliers.

The Initial EDA

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?

The EDA notes variables with skewed distributions very well. It also shows various transformations possible and the final selection nicely.

We would like to see boxplot to get a sense of spread and outliers for each variable.

Correlation wrt crime rate (R) is useful to show: e.g. offense mix has negligible correlation with crime rate. Should we still consider it in our model, and if so, why?

We would like the EDA to note influential outliers and possibly eliminate them if they happen to be due to a measurement error, for example.

Along with graphs, we would appreciate "points of interest" for each variable: what should we make note of this graph? What stands out about it? If there is nothing remarkable, it is still worth mentioning it.

EDA shows a correlation heat map, but does not try to analyze or provide remarks on anything interesting in the figure. The heatmap does not use the transformed variables, so the heatmap becomes less effective.

There isn't much description around *wage* variable and its transformation, so it is hard to follow what is being done with 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? 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?

Yes, available transformations and selection are shown well. Linear relationships are shown via scatterplots.

We especially liked the checks on assumptions of the CLR model.

Model 1 simply includes everything and is not very interesting in itself. Model 2 includes fewer variables; model 3 has even fewer. The fit is good for all three of these, but merits a question around transformations: were they necessary? Models 1 and 2 show influential outliers beyond Cook's distance 1.

We would appreciate why the model chose to use those variables: e.g., "We think a large number of young males can indicate aggressiveness and therefore we like to include it in the model". We did not understand why wages were included in the model.

Only Model 4 uses transforms and gets a fit of 0.49 per R^2. We think this is low, and it may be because the study did not remove outliers.

In general, we felt lack of enough explanation in text to understand the meaning of each visualization

Violation of instructions:

1) The team used ols_step_forward_p to find the the model with the best fit. ols_step_forward_p has not been covered in the course.
2) Concepts like bptest, shapiro.test, histogram of residuals were not covered in the course at the time of this draft.

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?

It is easy to find key coefficients in the regression table.

We did not see a lot of textual analysis of why certain variables were chosen.

We did not see practical significance for key effects either.

Violation of instructions:

1) The team used vcovHC for handling heteroskedasticity which was not covered in the course at the time of this draft.

The Omitted Variables Discussion

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?

There is a section on omitted variables, but a few variables are such that only a guess on the direction of bias can be made. There is no discussion on proxy variables.

It would be nice to have a list of how some of the omitted variables can bias some of the independent variables. e.g., education can negatively influence prbarr but positively influence taxpc.

Conclusion

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?

We do not see a cogent conclusion or a clear set of recommendations to the local government; this weakens the introduction statement "relationships are better understood we will use them to provide policy recommendations" . It can be useful to talk in terms of variables (polpc, prbarr, prbconv, density), but it is more useful to go holistic and talk in terms of what makes sense and what does not.

How would the report address questions such as: since police in high crime is intuitive, what should be the policy recommendation? An idea would be to check on the effectiveness of the police force. Given the low correlations of some of the variables with crime rate, can we have some policy changes (e.g. see avgsen)?

We can argue that 0.05 and 0.02 are not significant coefficients: they are very close to zero.

Persuasion and Impact

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

We think the conclusion is not as strong as we expected, based on the introduction. We would like a discussion of each variable from a policy perspective, along with the EDA: e.g., we found it surprising that probability of prison had very little correlation with crime rate.

The text to whitespace/R ratio is quite high in this report. What's more valuable is to show the analysis and understanding, while using the figures and formulas as props. The total page length requirement (of 20 pages) wasn't adhered to in this report.

A cosmetic but important glitch is the formatting of the report. It is hard to read it as a PDF perhaps because it was meant to be used as a notebook. At times, the sections form a continuous run due to missing formatting.