Lab 3 – Part 2 (Peer Review / Feedback)

MIDS Fall 2019 / w203 Tuesday 4pm

Group Name: Wu / Belfer / Beohar

Providing feedback for: Sayan Das / James Gao / Jacob Tosh

1.0 Introduction. Is the introduction clear? Is the research question specific and well defined? Could the research question lead to an actionable policy recommendation? Does it motivate the analysis? Note that we're not necessarily expecting a long introduction. Even a single paragraph is probably enough for most reports.

* Yes, the introduction is clear and explains a research question for which there is an obvious potential policy recommendation.

[vb ->] The research question is broken into two sub-parts and explores whether increasing police per capita have any impact on crime rates. It also explores the impact of tax revenue on police per capita and/or crime rates. This research question is clearly explained and in-line with expected analysis presented in the subsequent analysis. Ideally, it would have been better if the research

RB – Yes, the questions are very to the point and could lead to a clear recommendation. The data given can be used to answer these questions. The only comment would be to maybe justify the variables chosen from a cause vs effect perspective.

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?

* The team showed thorough understanding of the data by identifying anomalous outliers related to the prbarr (arrests per offense) and prbconv (convictions per arrest) columns. They justified the removal of those outliers. The use of the scatterplots were able to identify positive correlations as well as the potential issue of multicollinearity.

* As an added step, they could have used a correlation matrix to quantify the degree of correlation and risk for multicollinearity.

RB – The team noticed values over 1 in the arrest probabilities, and reasoned as to why it was anomalous. The scatterplot may not have been necessary if the correlation plot mentioned above was used instead.

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 at least two key explanatory variables and perform a thorough univariate analysis of each? 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 interperability? Are transformations used to expose linear relationships in scatterplots?

Lab 3 – Part 2 (Peer Review / Feedback)

MIDS Fall 2019 / w203 Tuesday 4pm

Group Name: Wu / Belfer / Beohar

Providing feedback for: Sayan Das / James Gao / Jacob Tosh

Is there enough explanation in the text to understand the meaning of each visualization?

* Yes - each step in the model building process is support by EDA. The outcome variable is appropriate, and the team identified at least two explanatory variables and explained them clearly. Transformations were taken carefully and explained.

4.0 Regression Models: Base Model. Does this model only include key explanatory variables? Does the team identify what they want to measure with each coefficient? Does the team interpret the result of the regression in a thorough and convincing manner. Does the team evaluate all 6 CLM assumptions? Are the conclusions they draw based on this evaluation appropriate? Did the team interpret the results in terms of their research question?

* Yes, the model only includes key explanatory variables, and the team explained what they were trying to measure with each coefficient. Interpretation of the coefficient may be modifed as per the comment below. The team indicated that they will evaluate the CLM assumptions in part 3. Additional discussion regarding the interpretation may be necessary - current interpretation seems to imply a causative effect between crime and police, while logically, it's only associative. A discussion of the omitted variables that might be underlying the associative effect would be useful at this stage.

* The sentence describing "the crime rate variable increases by 20 each time polpc increases by one" can be modified for easier interpretability. Since polpc is the # of police per capita, it's slightly unrealistic to increase the police presence to more than 1 officer per citizen. Perhaps easier to interpret with, "if there were an average of 0.1 officers per citizen, if police presence were increase by 0.1 so that there were 0.2 officers per citizen, crimes committed by each citizen would increase by 0.1 * 19.82 [coefficient of model] = 1.982".

4.1 Regression Model: Second Model. Does this model include covariates meant to increase the accuracy of the regression? Has the team justified inclusion of each of these additional variables? Does the team identify what they want to measure with each coefficient? Does the team interpret the result of the regression in a thorough and convincing manner. Does the team evaluate all 6 CLM assumptions? Are the conclusions they draw based on this evaluation appropriate? Did the team interpret the results in terms of their research question?

* The team includes covariates meant to increase the accuracy of the regression and justifies the inclusion of the additional variables as well as identifies what they want to measure. Interpretation of the results will likely change with the p-values/standard

error results, as coefficient size as a measure of effect size is only apples-to-apples when the features are similarly scaled. The team does tie the results to the ultimate research questions/policy recommendations (i.e., little effect on crime).

* Interpretation of the taxpc and polpc coefficients may need to change - taxpc coefficients will typically always be very small compared to polpc coefficients because average value of taxpc is 38 while average value of polpc is 0.0017022. Doesn't mean that taxpc has a small effect, it can also mean that the taxpc feature values tend to be very large compared to the other features.

* Adjusted R-squared may be a better outcome to look at as it is less sensitive to additional variables bumping up the R2 value.

4.2 Regression Model: Third Model. Has the team explained what value can be derived from this model? Does the team interpret the result of the regression in a thorough and convincing manner. Does the team evaluate all 6 CLM assumptions? Are the conclusions they draw based on this evaluation appropriate? Did the team interpret the results in terms of their research question?

* Similar to the above, the size of the coefficients is an indicator of how important a feature is only if the features are on similar scales.

* Adjusted R-squared may be a better outcome to look at as it is less sensitive to additional variables bumping up the R2 value.

RB – Could add a part about why those particular wage variables were chosen instead of other wage ones

4.3 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?

* Yes, the table is easily read, and the team mentions that practical significance will be discussed in part 3.

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?

* The report identified major sources of omitted variable bias. Estimate directions of bias were correct and explanations were clear. Would have liked to see

more discussion about which direction the omitted variable would have likely pushed each of the existing feature coefficients.

6.0 Conclusion. Does the conclusion address the high-level concerns of a political campaign? Does it raise interesting points beyond numerical estimates? Does it place relevant context around the results?

* The naive interpretation of the results would indicate that adding more police would increase crime, which is contrary to the goal of our policies (to reduce crime). Additional discussion to help the reader make the leap from this naive interpretation of the results to the writers' final recommendation to increase rather than decrease police presence would be helpful.

[vb ->] As mentioned above, the current interpretation seems to imply a causative effect between crime and police, while logically, it's only associative. However, the final interpretation and actionable recommendation seems to suggest otherwise. In current circumstances, police presence might be higher in areas that have higher crime rates. This does not mean increasing the police would lower the crime rates. There might be other external factors such as poverty, unemployment that might need to be tackled in order to fight higher crime rates.

As a political campaign, suggesting action be taken on those other “external” factors might make more sense than asking for greater police presence.

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?

* Main elements that would benefit from additional discussion are 1) size of coefficients vs size of features vs effect size of features 2) omitted variables and their potential effect on the feature coefficients [pushing higher or lower] 3) logical leap in conclusion regarding the naive interpretation of the results vs the final recommendation.