Lab 3-Draft feedback

For Huang, Lim, Wu, and Wu

From Essa, Lin, and Wheeler

1.0 Introduction. Is the introduction clear? Is the research question specific and well defined? Could the research question lead to an actionable policy reccomendation? 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 very clear. Great job.
- The research question: **"How can we leverage tax, policing, and housing policy in our campaign to lower crime?"** is specific and well-defined but we are not sure how you will answer it with statistics.
- Yes, the research question could lead to actionable policy reccomendations. For example, taxe rates, police per capita, and housing zones could be changes if the results motivat such changes.
- Yes the research question motivates the analysis. However, we do not understand why you chose taxes, police, and housing as predictors. Is it not possible that other variables are driving crime?

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?

- Yes, the team deleted the duplicate occurrences of county 193. Good job. Very sleek.
- Yes. you removed NA values & duplicates for county 193. Both are justified. Why did you not remove *year* since it is a constant?
- They were correct to (i) report that some *prbconv* > 1 and (ii) not remove these values since these 'probability' variables do not represent true probabilities. Also, they note that the wage data were likely censored (e.g. top and bottom coded).
- Yes, the report demonstrates a thorough understanding of the data. We are just not sure why you choosing the tax, police, and housing data as predictors- why not use model selection to inform what variables drive crime rather than impose your worldview onto the data?
- Note: please directly discuss all figures you present.
- Your scatterplot matrix with red-blue dots that depict collinearity is great. However, note, that only *perfect* collinearity is troublesome, as you mention later. Unperfect/weak collinearity is expected, as noted by Wooldridge. For this reason, we do not think you need to transform the wage data *for that reason*- maybe for other reasons...

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

- Yes, but it is still unclear why these predictors are choose *a priori* rather than using the model building process to *reveal* important predictors.
- Yes, very thorough univariate analysis was completed for the *crmrte*. The authors just need to be sure to mention each graph explicitly. We recommend that you label each figure with (figure 1,2...n) and cite them as such so you do not loose points.
- Yes, the team performed thorough univariate analysis on *taxpc*, *prbarr*, *wmed*, and *density*.
- Yes, they clearly explained why they choose these variables.
- Yes, these variables do reflect their research question.
- No, transformations were not included.
- No, transformatins were not used to reveal linearity in scatter plots.
- No, not every graph was mentioned. Please cite and discuss each graph.

Finally,

- Note that histograms and boxplots both show the same thing in different ways- you might consider just picking one to limit the total number of figures you need to discuss.
- Be careful with the **data linearity** assumption. Do you mean the model is **linear in the parameters**?
- Be careful to differentiate between the population error, u, and the residuals.
- Have you considered interaction effects?
- Have you considered transformations?

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 base model included key explanatory variables (taxpc, prbarr, density) that they intend to analyze.
- Yes, the team provided explaination of measurement of each coefficient.
- Yes, the team provided proper interpret the result of the regression. It will be more interesting to discuss the meanings of relationship among variables since they show correlations (taxpc and density has positive correlation; prbarr and density has negative correlation).
- Yes, the team clearly evaluate all 6 CLM assumptions in 4.0.2.
- Yes, the conclusions were clear and appropriate.
- Yes, the interpretation of results and conclusions of based model reflected their research question. Moreover, they made policy recommmendations based on each key explanatory variables.

- Additional comments:
  - The taxpc, prbarr, and density variables show strong positive skew, log-transforamtion may help on normal distribution.

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?

- Yes, the pctmin80, pctymle, prbconv, and polpc were introduced to imporve the accuracy of the regression from model 1.
- Yes, the team made justifications on including these variables. The team also stated that adding these variables was to increase the accuracy but not follow the research question.
- No, the purposes of adding these variables regarding to political campaign were weak.
- Yes, the team interpreted the result of the regression in 4.1.4 and evaluated all 6 CLM assumptions in 4.1.2.
- Yes, the conclusion had clear statistical interpretation on both original and adding variables.
- No, the interpretation of results did not clearly describe how to assist the policy making based on the adding variables.

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?

- Yes, the team explained the values can be derived from this model in 4.3.1.
- Yes, the team interpreted the result of the regression in 4.3.2 and evaluated all 6 CLM assumptions in 4.2.2.
- No, the conclusion mainly explained the statistical meanings with weak arguments toward research question and political campaign.
- No, the team did not interpret the results in terms of their research question based on the third model.

Additional suggestions:

- It might be more clear to have strong link among three model. In current version, the first model had the best arguments regarding to research question and political campaign. However, the conclusions and policy recommmendations based on model 2 and model 3 are weak.

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 model specifications were properly chosen.
- Yes, key coefficients were easy to find.
- Yes, practical significance for key effects were discussed in 4.3.1.

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 team did a good job of identifying sources of omitted variable bias and estimating the direction of bias.
- The rationale for each omitted variable and their potential bias is well reasoned and they have identified the practical implication of each.
- As far as missing any major sources of omitted variable bias, the team has covered what we also believe to be relevant and a few more as well. Of the sources they cite, "Social status (lower / middle / upper class)" is perhaps a little redundant since this variable can likely be very correlated to wages and job opportunities, which in turn are mentioned to account for some of the bias.

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 conclusion is presents the high level policy recommendations concisely. The details presented in the various sections go through a thorough discussion of the numerical aspect and practical steps that can be taken to reduce crime rate. However, It would be nice if they could recap the policy change suggestions in slightly more detail in the end again.

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?

- Overall the report is well organized and has pertinent data. One thing we would caution against is that since we are dealing with a snapshot of cross sectional data, the use of the word "true" in the covariate discussion may not be techincally accurate. e.g. 'Young males are also known to have a tendency in crime participation, hence, adding this as covariate could better identify the true effect size of the explanatory variables.'

In [ ]: