Lab 3 Feedback for Essa, Lin, and Wheeler

- **From: Ellie Huang, Kai Qi Lim, Justin Wu, Youzhi (Chloe) Wu**
- **W203: Statistics for Data Science, Spring 2019**
- **Section Number: 05**

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.

> The introduction is clear but the research question needs to be more specific. The introduction specifies studying crime rate to understand if it is possibly explained by certain variables. However, specifically identifying crime rate, a variable in the given data set, in the introduction may be too specific. It may be better discussed and justified in the section for model building.
>
> Also, we think the overarching theme is to apply this research to other areas, not North Carolina specifically. If this is how the team interpret it (and justify why it should be limited to North Carolina), then more elaboration is needed. Therefore, the introduction appears limiting when specifying North Carolina without explaining the context -- the dataset is taken from a subset of crime data in North Carolina. We are not sure the multicollinearity explanation for including wage variables make sense…? Otherwise, the layout of the objectives was good and clear.

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 noticed anomalies in the original dataset, and performed detailed data cleaning before their analysis. For abnormal data that are not removed (such as prbarr and prconv with values greater than 1), they provided reasonable justifications as to why not to remove, great job on that. Overall, the report demonstrates a thorough understanding of the data.

A few minor points to consider:

- Explain why county, west, central and urban were converted from numeric to factors
- Explain why prbconv was converted from factor to numeric
- Check for any duplicate records
- Perhaps more detail as to why everything was log transformed as opposed to a blanket random transformation method.
- It seems like multicollinearity was examined based of crime vs. demographics, crime, vs geography, and crime vs. economics, because there are 3 separate models to test these larger categories.
    - Why was polpc and mix included in the probabilities correlation plots -- because we are examining all probabilities?
    - To be safe it may be helpful to include why these categories were examined separately from each other (ie. demographics and economics if certain demographics make more money, or geography and economics if certain areas make more money). Might be something like we are looking for casual predictors among these 3 domains/categories so that's why we chose to examine each domain separately.

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 interoperability? Are transformations used to expose linear relationships in scatterplots? Is there enough explanation in the text to understand the meaning of each visualization?

The team's model building process is supported by thorough EDA.

- They selected crmrte as their outcome variable, which is considered as appropriate. They plotted histogram to perform univariate analysis of this outcome variable.
- The team's approach is different in that they run regression test on each of the variable and then select those that would significantly affect crmrte as key variables. However, before running the regression, there may need to be (1) more EDA on these key variables, (2) explanation on why selecting these key variables as to how they would address the research question, and (3) why certain transformations were performed and how they could be interpreted in a practical meaning.

A few points to consider:

- The comment "all the analysis were based on the log-transformed data" after the first histogram plots of all variables seemed out of place. It looked like the log transformation happened after that comment
- The inferential analysis section was not clear at first in explaining it was selecting for a few key explanatory models for Model #1 (base model)
- Perhaps emphasize the importance of the new selection method (individually testing each predictor against crmrate) earlier on so that the reader has a better idea of what to expect as the final modeling approach (building models for demographics, economics and geography and the weak linear relationship is what helped you come to the conclusion that each field needed to be individually tested so I would mention that earlier!). Helps reader know what to expect for in terms of base model, 2nd and 3rd model (in the parts below)

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?

- Comments on strong interpretations of statistical and practical significance could be added, using the numbers observed to inform the practicality.
- All CLM assumptions were covered.
- Perhaps some comments may be made on Cook's distance plot about data outliers, which was presented but not discussed.

It looks like the team decided to combined all the modeling sections below (model 1, 2, and 3) into one big model section, which was a bit confusing to read through at first. I was expecting to read model 1 with key variables first, then with other models adding in other covariates. This is a formatting feedback only. Overall the model diagnostics, regression tables, and explanations were clear.

Proposed formatting (to better explain my thoughts):

- Model 1 with key explanatory variables, diagnostic plots, regression table + explanations
- Model 2 with key explanatory variables + covariates, diagnostic plots, regression table + explanations
- Model 3 with all selected variables, diagnostic plots, regression table + explanations

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 identified what they want to measure with each coefficient. This should be m_associated? It is not clear which are the covariates, why they are included and that they are meant to increase the accuracy of the regression. The adjusted r.squared is actually lower than m_key which has less 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?

> The value of the model could be better explained, but the results of the regression and coefficients were thoroughly covered.

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?

> - It was easy to find key coefficients in the regression table.
> - However, practical significance for key effects was not discussed.

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 listed 5 categories of omitted variables and briefly explained whether the category/variable under/overestimate the impact on crime rates. There could be more explanation on estimating direction of 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?

> In stating that m_key is the most parsimonious model, it may be helpful to include AIC numbers for justification.
>
> The team clearly explained which model they recommend in the conclusion of the research paper. However, a final summary of the specific policy recommendations related to the selected key variables would have been helpful to tie back to the original context of this research paper. Otherwise, the team did a good job bringing up reproducibility and economic perspectives of why their selected model was a better choice.
>
> Lastly, for formatting change I would separate // make a new Conclusion section more distinct from the omitted variable section so that the reader has a better sense of what can and cannot be concluded from the data/information provided.

7.0 Other errors, faulty logic, unclear or not persuasive writing, or other less convinced elements

> Called out in each section feedback above instead of listing separately here.