### Proposed New Lab 2 Structure:

* Google Slides Link: <https://docs.google.com/presentation/d/1PqtuL0Z_C_VbpwO6OY9uYU3Fcy5dJkuS7LoKiqDDzQM/edit#slide=id.g85fb292069_1_82>
* Introduction
  + <Your introduction should present a research question and explain the concept that you’re attempting to measure and how it will be operationalized. This section should pave the way for the body of the report, preparing the reader to understand why the models are constructed the way that they are. It is not enough to simply say “We are looking for policies that help against COVID” Your introduction must do work for you, focusing the reader on a specific measurement goal, making them care about it, and propelling the narrative forward. This is also good time to put your work into context, discuss cross-cutting issues, and assess the overall appropriateness of the data.>
  + **Rubric:** Is the introduction clear? Is the research question specific and well defined? Does the introduction motivate a specific concept to be measured and explain how it will be operationalized. Does it do a good job of preparing the reader to understand the model specifications?
  + 6 Sections:
    - 1. Background on your research topic
    - 2. Your Research question
    - 3. Very detailed operationalization
    - 4. Describe why did you pick the variables you picked and how are you using them in your analysis
    - 5. Describe why you **didnt** use other variables.
    - 6. Describe what variables you omitted.
  + Operationalization - Initial Data Loading and Cleaning
    - **Rubric:** 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?
    - **No Graph or Data Dumps**: Only include graphs (histograms) if it is additive to the report/if we have a sentence or two to explain it. Graphs without explanations = Delete.
    - For SEs: Use Homoskedastic SEs. Do not use Robust SE’s because they are Heteroskedastic.
    - Top coding and bottom coding means: Elaborate about decisions to remove or keep outliers (NY) or filling in garbage/non-ending data (dates)
  + Operationalization - The Model Building Process.
    - <You will next build a set of models to investigate your research question, documenting your decisions. Here are some things to keep in mind during your model building process:
      * What do you want to measure? Make sure you identify one, or a few, variables that will allow you to derive conclusions relevant to your research question, and include those variables in all model specifications.
      * Is your modeling goal one of description or explanation?
      * What covariates help you achieve your modeling goals? What covariates are problematic, either due to collinearity, or because they are outcomes that will absorb some of a causal effect you want to measure?
      * What transformations, if any, should you apply to each variable? These transformations might reveal linearities in scatterplots, make your results relevant, or help you meet model assumptions.
      * Are your choices supported by exploratory data analysis (EDA)? You will likely start with some general EDA to detect anomalies (missing values, top-coded variables, etc.). From then on, your EDA should be interspersed with your model building. Use visual tools to guide your decisions. You can also leverage statistical tests to help assess whether variables, or groups of variables, are improving model fit.>
    - <At the same time, it is important to remember that you are not trying to create one perfect model. You will create several specifications, giving the reader a sense of how robust (or sensitive) your results are to modeling choices, and to show that you’re not just cherry-picking the specification that leads to the largest effects.>
    - <Guided by your background knowledge and your EDA, other specifications may make sense. You are trying to choose points that encircle the space of reasonable modeling choices, to give an overall understanding of how these choices impact results.>
    - **Rubric**: 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 one, or very small number of explanatory variables and perform a thorough univariate analysis of each one? 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 interpretability? Are transformations used to expose linear relationships in scatterplots? Is there enough explanation in the text to understand the meaning of each visualization?
* Model 1:
  + <One model with *only the key variables* you want to measure (possibly transformed, as determined by your EDA), and no other covariates (or perhaps one, or at most two, covariates if they are so crucial that it would be unreasonable to omit them)>
  + **Rubric:** Model 1 Does this model only include key explanatory variables? Do the variables make sense given the measurement goals? Did the team apply reasonable transformations to these variables, to capture the nature of the relationships?
  + Address all CLM assumptions
    - 1 Linear in parameters
    - 2 Random iid sampling
    - 3 No perfect multicollinearity
    - 4 Zero conditional mean (Exogeneity, linear conditional expectation)
    - 5 Homoskedasticity
    - 6 Normality of Error Terms
    - <As a team, evaluate all of the CLM assumptions that must hold for your models. However, do not report an exhaustive examination all 5 CLM assumption. Instead, bring forward only those assumptions that you think pose significant problems for your analysis. For each problem that you identify, describe the statistical consequences. If you are able to identify any strategies to mitigate the consequences, explain these strategies. Note that you may need to change your model specifications in response to violations of the CLM.>
    - **Rubric:** Has the team assessed each of the CLM assumptions (including random sampling)? Did they use visual tools or statistical tests, as appropriate? Did they respond appropriately to any violations?
* Model 2:
  + <One model that includes key explanatory variables and covariates that you believe advance your modeling goals without introducing too much collinearity or causing other issues. This model should strike a balance between accuracy and parsimony and reflect your best understanding of the relationships among key variables.>
  + **Rubric:** Does this model represent a balanced approach, including variables that advance modeling goals without causing major issues? Does the model succeed in reducing standard errors of the key variables compared to the base model? Does it capture major non-linearities in the joint distribution of the variables?
  + Address all CLM assumptions
    - 1 Linear in parameters
    - 2 Random iid sampling
    - 3 No perfect multicollinearity
    - 4 Zero conditional mean (Exogeneity, linear conditional expectation)
    - 5 Homoskedasticity
    - 6 Normality of Error Terms
    - <As a team, evaluate all of the CLM assumptions that must hold for your models. However, do not report an exhaustive examination all 5 CLM assumption. Instead, bring forward only those assumptions that you think pose significant problems for your analysis. For each problem that you identify, describe the statistical consequences. If you are able to identify any strategies to mitigate the consequences, explain these strategies. Note that you may need to change your model specifications in response to violations of the CLM.>
    - **Rubric:** Has the team assessed each of the CLM assumptions (including random sampling)? Did they use visual tools or statistical tests, as appropriate? Did they respond appropriately to any violations?
* Model 3:
  + <One model that includes the previous covariates, and many other covariates, erring on the side of inclusion. A key purpose of this model is to demonstrate the robustness of your results to model specification. (However, you should still not include variables that are clearly unreasonable. For example, don’t include outcome variables that will absorb some of the causal effect you are interested in measuring.)>
  + **Rubric**: Does this model represent a maximalist approach, erring on the side of including most variables? Is it still a reasonable model? Are there any variables that are outcomes, and should therefore still be excluded? Is there too much multicollinearity, to the point that the key causal effects cannot be measured?
  + Address all CLM assumptions
    - 1 Linear in parameters
    - 2 Random iid sampling
    - 3 No perfect multicollinearity
    - 4 Zero conditional mean (Exogeneity, linear conditional expectation)
    - 5 Homoskedasticity
    - 6 Normality of Error Terms
    - <As a team, evaluate all of the CLM assumptions that must hold for your models. However, do not report an exhaustive examination all 5 CLM assumption. Instead, bring forward only those assumptions that you think pose significant problems for your analysis. For each problem that you identify, describe the statistical consequences. If you are able to identify any strategies to mitigate the consequences, explain these strategies. Note that you may need to change your model specifications in response to violations of the CLM.>
    - **Rubric:** Has the team assessed each of the CLM assumptions (including random sampling)? Did they use visual tools or statistical tests, as appropriate? Did they respond appropriately to any violations?
* A Regression Table (Stargazer)
  + <You should display all of your model specifications in a regression table, using a package like stargazer to format your output. It should be easy for the reader to find the coefficients that represent key effects near the top of the regression table, and scan horizontally to see how they change from specification to specification. Make sure that you display the most appropriate standard errors in your table, along with significance stars. In your text, comment on both statistical significance and practical significance. You may want to include statistical tests besides the standard t-tests for regression coefficients.>
  + **Rubric**: 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?
  + **Rubric**: Do the plots, figures and tables that the team has chosen to include successfully move forward the argument that they are making? Do they have a good ratio of “Information to Ink” (Tufte)? Has the team chosen the most effective method (a table or a chart) to display their evidence? Is that table or chart as communicative as it can be? Is every single plot, figure, or table that is included in the report referenced in the main text?
* Omitted Variables Discussion
  + <If your team has chosen an explanatory (i.e. causal) question to evaluate, then identify what you think are the 5 most important omitted variables that bias the coefficients you care about. For each variable, you should reason about the direction of bias caused by omitting this variable. If you can argue whether the bias is large or small, that is even better. State whether you have any variables available that may proxy (even imperfectly) for the omitted variable. Pay particular attention to whether each omitted variable bias is towards zero or away from zero. You will use this information to judge whether the effects you find are likely to be real, or whether they might be entirely an artifact of omitted variable bias.>
  + **Rubric**: 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?
* Conclusion
  + <Make sure that you end your report with a discussion that distills key-takeaways from your estimates, addresses your research question, and draws attention to larger contexts.>
  + **Rubric:** Does the conclusion address the research question? Does it raise interesting points beyond numerical estimates? Does it place relevant context around the results? Are there any other errors, faulty logic, unclear or unpersuasive writing, or other elements that leave you less convinced by the conclusions?
* Other internal notes:
  + 8000 word limit
  + Need to create a presentation slide

### 

### 

### 

### Potential Questions

* Missing info: Covid data **over time**
* Are masks effective in preventing COVID? (4 mask columns vs. log(COVID-19 cases) to get a non-percent I believe)
* Does poverty correlate to a higher COVID infection rate? (% living under federal poverty line)
* Do people with more access to money correlate to a lower COVID infection rate? (median household income, a high “weekly UI maximum amt”)
* Do minorities get infected with COVID more often than non-minorities? (all races vs. white)
* Does a high number of homeless people correlate with a higher COVID infection rate (number homeless)?
* Does a high Medicaid expenditure translate to lower COVID-related deaths? (Medicaid Expenditures vs. all-cause deaths)
* **Is patient race a larger contributor to mortality rate than state policies/preparedness?**
  + Factors that could drive mortality rate
    - Priority 1: Race (Black + Latinx + Other vs White, and each as a standalone race)
    - Priority 1: Mask Adherence
    - Priority 2: Preparedness (Hospital Beds, ICU Beds, # of Hospitals)
    - Priority 3: Increase in park mobility
    - *Priority 3: Percent living under the federal poverty line (2018)*
    - *Priority 3: Possibly Age*
  + Conclusion: Assuming significance, a Higher coefficient for patient race vs. state policies will prove our hypothesis correct, and we will reject the null hypothesis. Else fail to reject null.
* Not part of the model, but look at the proportion of number infected by the population of the various racial groups

**Mark OH Notes:**

* **You're not going to solve all the inadequacies of this data set, but it's important to note where some of these inadequacies lie.**
* He likes our research q, but he says his main focus will not be on the research q, but on how we operationalize the variables
  + Problem with race is from a causal perspective: you are born a particular ethnicity, so there is no randomization across race. So this is an issue when you are creating causal models. **Non randomizable element in the model**. Controlling for it is going to be very important.
  + Group races as option: nonwhite vs. white. Nothing necessarily wrong with grouping stuff.
  + Age: Can break down into 2-3 categories. Can reduce # of predictors.
  + Controlling for race is going to be the main issue.
  + **Mortality rate (a 0 to 1 number): lots of omitted variables.** Do a qqnorm plot on the residuals. How do we account for more than one omitted variable?
  + If many variables that are insignificant = high level of collinearity between the control variables.
  + “Don't want someone giving me 20 variables on a 50 sample data set.” Limited # of samples, cross validation approach (K-fold approach) if very few samples.
* **Dont** compare adjusted R2 between mortality\_rate vs. log(mortality\_rate) models!
* Mark Rule of thumb: Only have 1/10th to 1/5th the samples: **5-8 variables = max, if there's only 50 states. Don’t** go beyond quadratic at all in fitting anything. Avoid cubic.
* States are **not** iid. Reasons: Geography, commercial practices are different, political nature. If you wanted to avoid sales tax/alcohol, you would go to new hampshire to buy things as opposed to massachusetts. Commercial relationships are diff so that prevents iid. Practices in rural areas (red) are very different from pacific portions of the country (blue).
* When you sample from a population, you choose some people and not others. Did this choice align themselves with the outcome you have, or mechanical in nature?
* 6 assumptions: iid, linear, 0 conditional mean, heteroskedasticity, normality of errors, no perfect collinearity

---

title: "Lab 2 DRAFT: Regression to Study the Spread of Covid-19"

author: 'Kevin Fu, Jamie Smith, Ratan Singh'

date: "11/13/2020"

output: pdf\_document

---

```{r, message=FALSE}

knitr::opts\_chunk$set(echo = TRUE)

if(!require(stargazer)){install.packages("stargazer")}

if(!require(readxl)){install.packages("readxl")}

library(sandwich)

library(tidyverse)

library(stargazer)

library(readxl)

data <- read\_excel("covid-19.xlsx", "Covid-19")

head(data)

```

**### 1. An Introduction**

*Your introduction should present a research question and explain the concept that you're attempting to measure and how it will be operationalized. This section should pave the way for the body of the report, preparing the reader to understand why the models are constructed the way that they are. It is not enough to simply say "We are looking for policies that help against COVID" Your introduction must do work for you, focusing the reader on a specific measurement goal, making them care about it, and propelling the narrative forward. This is also a good time to put your work into context, discuss cross-cutting issues, and assess the overall appropriateness of the data.*

The amorphous nature of Covid-19 has baffled many medical practitioners. The virus has turned the world upside down and has continued to do so throughout 2020. Though the impacts of this virus have been universally felt, our research team postulates that the virus impacts certain demographic groups more acutely than others. The team’s hypothesis is that patient race is a larger contributor to mortality rate than statewide mask policies and preparedness.

To measure these factors, we will be looking at the mortality rate across the 50 states for White, Hispanic, and Black patients. We’ll then review variables related to government mask mandates, and the number of hospital and ICU beds available to patients.

We will run a regression analysis to determine if factors related to the racial demographic prove to be better predictors of mortality rate than factors related to government mask mandates and hospital beds.

* Is patient race a larger contributor to mortality rate than state policies/preparedness?
* As there has been lots of discussion in different media outlets regarding the effectiveness of wearing masks, closing down the economy, as well as a higher COVID prevalence rate in minorities, we wanted to see if this was indeed the case by looking at these variables’ impact on mortality rate (our dependent variable).

**### 2. A Model Building Process**

You will next build a set of models to investigate your research question, documenting your decisions. Here are some things to keep in mind during your model building process:

1. \*What do you want to measure\*? Make sure you identify one, or a few, variables that will allow you to derive conclusions relevant to your research question, and include those variables in all model specifications.

- Factors that could drive mortality rate: 1) Race (Black + Latinx + Other vs White, and each as a standalone race), 2) Mask Adherence, 3) Preparedness (Hospital Beds, ICU Beds, # of Hospitals), 4) Increase in park mobility

- Other Factors that we will need to do further work on: 1) Percent living under the federal poverty line 2) Possibly Age 3) Number of Homeless

- Operationalization: COVID Deaths by race = # COVID Deaths by race/# COVID Cases by race

2. Is your modeling goal one of description or explanation?

- Descriptive model

3. What [covariates](https://en.wikipedia.org/wiki/Dependent\_and\_independent\_variables#Statistics\_synonyms) help you achieve your modeling goals? What covariates are problematic, either due to \*collinearity\*, or because they are outcomes that will absorb some of a causal effect you want to measure?

- As mentioned in 1), our covariates would include race, mask adherence, hospital preparedness, park mobility and % living under the poverty line (possibly age as well).

- The ICU and in-patient beds variable has some multicollinearity with other independent variables. The mask mandate also has some collinearity with other independent variables. However, since ICU beds have significance in our t-test (as we will see later), we wouldn't necessarily remove the variable outright. If the variable was insignificant, we would then discard it without hesitation.

4. What \*transformations\*, if any, should you apply to each variable? These transformations might reveal linearities in scatterplots, make your results relevant, or help you meet model assumptions.

- We quickly notice that DeathRate is a much higher number compared to other variables and taking a logarithmic value of DeathRate hence makes more sense.

- We then plot log(DeathRate) Vs various other variables.

- We notice a non-linear relationship between HispanicMortality and log(DeathRate) - hence we prefer to include a quadratic term for HispanicMortality in the regression equation.

- We also notice slight nonlinearity for the WhiteRaceMortality and log(DeathRate)- we would try to include a quadratic term for this as well.

5. Are your choices supported by exploratory data analysis (\*EDA\*)? You will likely start with some general EDA to \*detect anomalies\* (missing values, top-coded variables, etc.). From then on, your EDA should be interspersed with your model building. Use visual tools to \*guide\* your decisions. You can also leverage statistical \*tests\* to help assess whether variables, or groups of variables, are improving model fit.

At the same time, it is important to remember that you are not trying to create one perfect model. You will create several specifications, giving the reader a sense of how robust (or sensitive) your results are to modeling choices, and to show that you're not just cherry-picking the specification that leads to the largest effects.

- Our choices are supported by EDA.

- Anomalies:

- “NR”: Replaced with NA.

- “<0.01”: Replaced with 0.001. 0.001 is picked to represent lowest percentage value in the excel which is non-zero

- Mask Mandate dates have “0” in them - replaced with 10/30/2020. One of the dates mentioned is 12/31/99 - Bogus date, replaced by 10/30/2020. We picked 10/30/2020 since this was the latest date for the available data on the excel sheet.

At a minimum, you should include the following three specifications:

Create plot for ratio of Black Percentage, White Percentage, Hispanic Percentage and the death rates

\*\*Model 1\*\*: One model with \***only the key variables**\* you want to measure (possibly transformed, as determined by your EDA), and no other covariates (or perhaps one, or at most two, covariates if they are so crucial that it would be unreasonable to omit them)

Dependent: cvd$deathrate

Independent: cvd$whiteracedeath, cvd$blackracedeath, cvd$icubedspercvdpat

\*\*Model 2\*\*: One model that includes \***key explanatory variables and covariates** that you believe advance your modeling\* goals without introducing too much collinearity or causing other issues. This model should strike a balance between accuracy and parsimony and reflect your best understanding of the relationships among key variables.

Dependent: log(cvd$deathrate)

Independent: cvd$whiteracedeath, cvd$blackracedeath, cvd$hispanicracedeath, cvd$inpatbedspercvdpat, cvd$icubedspercvdpat, cvd$povertyperc

\*\*Model 3\*\*: One model that includes the \***previous covariates, and many other covariates\*, erring on the side of inclusion**. A key purpose of this model is to demonstrate the robustness of your results to model specification. (However, you should still not include variables that are clearly unreasonable. For example, don't include outcome variables that will absorb some of the causal effect you are interested in measuring.)

Dependent: log(cvd$deathrate)

Independent: cvd$whiteracedeath, cvd$blackracedeath, cvd$hispanicracedeath, cvd$inpatbedspercvdpat, cvd$icubedspercvdpat, cvd$povertyperc, cvd$numberofcases, cvd$parkmobility

Guided by your background knowledge and your EDA, other specifications may make sense. You are trying to choose points that encircle the space of reasonable modeling choices, to give an overall understanding of how these choices impact results.

**### 3. Limitations of your Model**

As a team, evaluate all of the CLM assumptions that must hold for your models. However, do not report an exhaustive examination of all 5 CLM assumptions. Instead, bring forward only those assumptions that you think pose significant problems for your analysis. For each problem that you identify, describe the statistical consequences. If you are able to identify any strategies to mitigate the consequences, explain these strategies.

Note that you may need to change your model specifications in response to violations of the CLM.

In order to better meet some of the CLM assumptions the team performed a number of data transformations on our variables of interest. The mortality rate by race was the first that we focused on transforming to better fit our CLM assumptions

In reviewing the data for the mortality rate for whites, we decided to square the values to allow for better spread. However, in doing this we start to see a bit of skew in the data which would ideally not be present. When evaluating the Black mortality rate, we performed a log transformation, which revealed a few potential outliers which we will test for to see if those data are unreliable and should be removed from the data set. Lastly, the hispanic rate required no transformation and had a good dispersion across the data.

For our measures of preparedness, we evaluated the number of ICU and Patient beds available in terms of the number of patients in the hospital. In reviewing this data, we found that there was a slight rightward skew in the data.

**### 4. A Regression Table**

You should display all of your model specifications in a regression table, using a package like [`stargazer`](https://cran.r-project.org/web/packages/stargazer/vignettes/stargazer.pdf) to format your output. It should be easy for the reader to find the coefficients that represent key effects near the top of the regression table, and scan horizontally to see how they change from specification to specification. Make sure that you display the most appropriate standard errors in your table, along with significance stars.

In your text, comment on both \*statistical significance and practical significance\*. You may want to include statistical tests besides the standard t-tests for regression coefficients. Mention vif.

**### 5. Discussion of Omitted Variables**

If your team has chosen an explanatory (i.e. causal) question to evaluate, then identify what you think are the 5 most important \*omitted variables\* that bias the coefficients you care about. For each variable, you should \*reason about the direction of bias\* caused by omitting this variable. If you can argue whether the bias is large or small, that is even better. State whether you have any variables available that may proxy (even imperfectly) for the omitted variable. Pay particular attention to whether each omitted variable bias is \*towards zero or away from zero\*. You will use this information to judge whether the effects you find are likely to be real, or whether they might be entirely an artifact of omitted variable bias.

- 1. "Mixed" Race variable, particularly for those that identify as both white and black

- Reasoning: In modern times, people have become more intertwined with each other. Over time, more of us will identify as more than one race. This blurs the data set; since there is no option to choose more than one race, we are assuming that respondents chose their "primary" demographic as opposed to checkmarking all of their identifying features. For example, our data may have bias if someone is both white and black, but responded "black" becuase they mostly identified as black instead.

- Direction of Bias (towards or away from 0): Towards 0

- Large or Small Effect: Small

- Any variables that may proxy for the omitted variable: The "Other Race" variable

- 2. Ages of all COVID-positive patients, particularly ages 50+

- Reasoning: We only have the demographic data by state. However, we don't have the demographic data of COVID-positive patients by state. We may have omitted this very important variable, especially for patients age 50+, since a person's age may very much influence mortality rate.

- Direction of Bias (towards or away from 0): Away from 0

- Large or Small Effect: Large

- Any variables that may proxy for the omitted variable: Age distribution in state, regardless of COVID infection

- 3. Access/affordability of COVID treatments

- Reasoning: In states where COVID is taken seriously and treatments are readily available and affordable, people are more likely to 1) discover that they have the disease, 2) have a uncrowded hospital to go to in the event they do not feel well, and 3) are able to afford the treatment. Not everyone can receive the same type of treatment and attention our President recently did. Therefore, low access/affordability to COVID treatments may have a relationship with a higher mortality rate.

- Direction of Bias (towards or away from 0): Away from 0

- Large or Small Effect: Large

- Any variables that may proxy for the omitted variable: Tests performed, Tests per 100K

- 4. isUrban or isRural boolean

- Reasoning: Since people are more tighly packed together in urban cities as opposed to people in parts of the rural US, this might be an omitted variable that has a significant relationship with mortality rate. Our hypothesis would be that people in the rural parts of the country would have a lower mortality rate vs. people in urban parts of the country.

- Direction of Bias (towards or away from 0): Away from 0

- Large or Small Effect: Large

- Any variables that may proxy for the omitted variable: Low Population density per square miles, Low median household income

- 5. \*\*Serious\*\* pre-existing health conditions

- Reasoning: Pre-existing health conditions definitely affect mortality rate. However, this is very hard to measure as this gets very granular. There is also a range of serious to non-serious pre-existing health conditions, and so we need to differentiate between those. Our hypothesis is that pre-existing serious health conditions may affect mortality rate after COVID infection.

- Direction of Bias (towards or away from 0): Towards 0

- Large or Small Effect: Large

- Any variables that may proxy for the omitted variable: Nonelderly Adults Who Have A Pre-Existing Condition

- 6. Availability/affordability of COVID testing

- Reasoning: In states where COVID is taken seriously and testing is widely available, people are more likely to get tested on a regular basis. Therefore, they may discover the disease earlier even when asymptomatic, and be able to get treatment earlier. In states without widely available/affordable COVID testing, people may go days/weeks without knowing they have the virus, and therefore delayed discovery and treatment may cause their unfortunate death in the future.

- Direction of Bias (towards or away from 0): Away from 0

- Large or Small Effect: Large

- Any variables that may proxy for the omitted variable: Tests performed, Tests per 100K

**### 6. Conclusion**

Make sure that you end your report with a discussion that distills key-takeaways from your estimates, addresses your research question, and draws attention to larger contexts.

- Key takeaways from our estimates:

- Address the research question:

- Attention for larger contexts:

### Submission

- Submit your draft report via ISVC; please do not submit via email.

- Submit 2 files:

1. A pdf, html or md file including the summary, the details of your analysis, and all the R codes used to produce the analysis. \*Please show code in your compiled file.\*

2. The Rmd or source file used to produce the pdf file.

- Only one group member needs to submit the files.

- Be sure to include the names of all team members in your report. Place the word 'draft' in the file names.

- Please limit your submission to 8000 words, excluding code cells and R output.