A Causal Research Investigation of COVID-19

Camilla Bendetti

Slide 1:

This project is a multi-modal planning of a causal research investigation. KG and I wanted to focus on the coronavirus outbreak because it is a topic surrounded by so much uncertainty at this time. We are hoping that this causal research investigation helps to answer questions surrounding the spread of COVID-19 in the United States. I will plan several different avenues of investigation for this study. I would like to thank KG Nguyen for her collaboration on this project and I would like to thank my professor, Leslie Myint, for her guidance.

Slide 2:

The first step is to define the research question of this investigation. The goal of this investigation is to estimate the effect of wearing masks on the chances of contracting COVID-19. I am examining the effect of wearing masks on the chances of contracting COVID-19.

Slide 3:

This question interested me because of the surge of calls to actions to wear masks during quarantine. While some state governments require the use of facemasks when out in public,

many more have yet to require this precaution. While the CDC now *recommends* that everyone wears a mask when they are out in public, there is still much debate about if masks are really effective. Many people choose not to wear masks in public because there is no clear information about whether or not wearing a mask will make a significant difference in the longevity of COVID-19 or the stay-at-home orders.

The image pictured on this slide is something that many of my Facebook friends have reposted, however, it is shared with little to no other context. This shows the uncertainty around wearing masks and how people so easily follow trends. Therefore, the goal of this project is to propose a causal research investigation of the effect of wearing masks on the spread of COVID-19.

Slide 4:

This slide shows the directed acyclic graph (DAG) that KG and I created, based on "expert knowledge", in order to represent different sources of concern for an observational study. Our selection node, C, is people in grocery stores. We choose to select people who shop in grocery stores because grocery stores are essential businesses, meaning they are open during quarantine, and places when people visit usually once a week. Because there are usually lines to enter a grocery store, as they are limiting the number of people who can be in the store at one time, it could be a good time to survey people while they are waiting in line. We would follow up with these people two weeks later to see if they had contracted the virus. C includes those people who were chosen to participate in our study and those who follow up with us afterward.

One factor that would play a role in if they responded 2 weeks later is their personality. We realize many people may not follow up afterward because they either do not care anymore or they want to keep their test results private. We understand that COVID-19 is a sensitive topic people deal within their own ways.

The treatment is wearing masks and the outcome is contracting COVID-19. Our mediator variable, which will be of interest for subsequent mediation analysis in step four, is handwashing. We assume that people who wear masks care more about their health and other people's health than people who don't wear masks. Therefore, if they tend to care more about their health they are probably washing their hands more frequently and taking other precautions to avoid contracting COVID-19, not just wearing a mask.

Other important variables key to performing a valid mediation analysis include confounding variables. The confounding variables influence both the outcome and the treatment. Our confounding variables include age, occupation, high-risk individuals, income, media awareness.

We chose age because people who are older are more likely to die from the COVID-19. Therefore, they would be more likely to wear masks to reduce their chances of contracting the disease and they are at a higher risk to contract the virus because they have weaker immune systems.

We choose occupation as a confounding variable because there are many people who are considered essential workers, like nurses, and are potentially more exposed to the virus than those of us who stay at home. Therefore they would be more likely to wear a mask and have a higher risk of contracting COVID-19.

The reason we chose high-risk individuals is similar to the reason for age as a confounding variable. Excluding age, high-risk individuals include but are not limited to people with lung problems, heart problems, and people who are immunocompromised.

We included income as a confounding variable because people who have a higher income usually have the ability to work from home or not work at all and therefore might not wear masks as much and have a lower chance of contracting the virus. Those with a lower income may be forced to work in order to sustain themselves and their families, requiring them to leave the house often, wear masks, and come in contact with more people who have COVID-19.

Finally, we added media awareness as a confounding variable. The more access people have to social media, television, and news the more they will hear about COVID-19 which will influence how they act.

Slide 5:

Step 3: IP Weighting Analysis

In Step Three, we will present an outline of an IP weighting analysis. Our research goal is to estimate the effect of wearing masks on the chances of contracting COVID-19 in the whole population.

First, we need to build a propensity score model. We will estimate the propensity scores by fitting a logistic regression model of wearing masks by the following variables. The variables that we included in our propensity score model include age, occupation, high-risk individuals,

income. We did not include media awareness in our propensity model because it is unmeasured. Unlike age and income, it is much more difficult to measure how aware someone is of what is going on in the media and how much media they consume. Age is quantitative, Occupation is categorical, high-risk individuals are categorical, and income is quantitative.

In order to ensure that our propensity score model is built well, we should consider that not all trends are linear and that non-linear trends might be warranted in our model or that we should consider including polynomial relationships. We could also add interaction terms to the model. Most importantly, we want the variables in the model to describe mask-wearing, the treatment, in an appropriate way. We can determine this by creating visualizations that show these relationships.

Slide 6:

In the IP weighting analysis above we are generalizing on people who respond with their test results two weeks later or the uncensored people in our study. However, in order to find the average causal effect we need to account for everyone and not just the shoppers who had been censored. We can account for this by creating a censoring probability model which looks at censoring given wearing masks and personality. Wearing masks is the treatment variable and personality d-separates the censoring variable and the outcome. Personality is a categorical variable. The final inverse probability weights result from the product of the IP treatment weights and the IP censoring weights. These final weights are used to upweight the treated

populations and the untreated populations in order to get comparable populations for the two.

Once we have these final weights, we can fit marginal structural models (MSMs) that will help us to estimate the average potential outcome, where the outcome is contracting COVID-19 when everyone is not wearing masks, and the average causal effect. The average causal effect is of interest here. It is the change in the average potential COVID-19 infection rate of a population if everyone was wearing masks compared to if no one was wearing masks.

Slide 7:

In step four, our goal is to present an outline for mediation analysis. The research question for the mediation analysis is “How much of the effect of wearing masks on chances of contracting COVID-19 is mediated by handwashing?”. We hypothesized that wearing masks may prompt some people to be more hygienic and care about other people's health, therefore they are more likely to frequently wash their hands than people who do not wear masks. Our mediator variable here is handwashing. Therefore, our DAG represents two causal pathways from the treatment to the outcome. There is a direct pathway from wearing masks to contracting COVID-19, representing that wearing masks acts directly on contracting the virus. There is also a causal pathway that includes the mediator variable, representing that the pathway from wearing masks to contracting COVID-19 includes handwashing as a middle node between the two.

The first step is to create a logistic regression model for the mediator. Our confounders include age, occupation, high-risk individuals, and income. Depending on what kind of variable

the mediator is, we need to fit an appropriate model. We can do this creating visualization to show the relationships between the mediator and the quantitative confounding variables, age, and income. The next step is to create a model for the outcome. We can create a logistic regression model for the outcome in the same way that we created the mediator model. In the outcome model, we will include the treatment variable, wearing masks, the mediator variable, and the confounding variables.

After we fit these models we would use the mediate( ) package to estimate the mediation effects. In a mediation analysis, there are four assumptions. The first assumption is that there is "no unmeasured confounding of the treatment-outcome relationship." The second assumption is that there is "no unmeasured confounding of the mediator-outcome relationship." The third assumption is that there is "no unmeasured confounding of the treatment-mediator relationship." The final assumption is that "no confounder of the mediator-outcome relationship is affected by treatment."[1] The mediate() function creates a summary of both of the models that we just made.

In order to identify the natural effects (NDE and NIE), all four of the assumptions need to hold. For the Natural Direct Effect (NDE) we fix the value of the mediator. In the mediate() output the NDE is represented by the ADE(control). So, we would set handwashing to be what it would have been when no one is wearing masks. In other words, we would set everyone to their normal hand washing routines. For the NDE we are looking for the average change in probability of contracting COVID-19 by comparing two scenarios. In the first scenario, we keep normal handwashing routines when wearing masks is not required but have everyone wear masks and in scenario two we have no one wear masks and keep normal handwashing routines as if wearing

[1] https://lmyint.github.io/causal_spring_2020/graphical-structure-of-mediation.html

masks were not required.

For the Natural Indirect Effect (NIE) part of the effect of mask-wearing is the unintended consequence of handwashing more or handwashing less. In the mediate() summary, the NIE is represented by the ACME(treated). For the NIE we are looking for the average change in the probability of contracting COVID-19 in two different scenarios. In the first scenario, we have everyone wear masks and let handwashing occur naturally as the result of wearing masks. In the second scenario we have everyone wear masks but let handwashing occur naturally as the result of no one wearing masks.

For both the NIE and NDE we can examine the 95% confidence intervals in order to look for evidence of a true indirect or direct effect of the broader population. If zero is contained in the confidence interval, that means that there is not a true causal effect. If we look at Prop. Mediated (treated) we can compute what fraction the NIE is of the total effect. This can help us quantify the effect of wearing masks on contracting COVID-19 that is through hand washing.

Slide 8:

In Step 5, our goal is to present a brief outline of how causal discovery can be used to check the sensitivity of our results to the expert knowledge encoded in our DAG. The results of our analysis will change if we change something in the original choices about what to do in the analysis. These changes can include something like adding unmeasured confounders or connecting confounders in a different way. The results of causal discovery give us equivalence classes. Equivalence classes show us how we can orient edges differently than we did using our

"expert knowledge." The results of causal discovery create a trickle-down effect that can cause us to change edges around in our original DAG which will lead us to new results.

For causal discovery, we will use the Spirtes-Glymour-Scheines (SGS) algorithm in order to get the equivalence class. We can use this equivalence class in conjunction with our knowledge and other information to uncover the true structure of the DAG. The first step is to start with an undirected version of the DAG we created that included the treatment and outcome as well as the confounders, age, occupation, high-risk individuals, and income. Next, we will do multiple rounds of testing to check whether or not the X and Y variables are conditionally independent given Z. This means that "For every pair of variables X and Y and all other sets Z of other variables, check if we can make $X \perp Y \mid Z$."[2] If X and Y are conditionally independent given Z, the next step will be to remove the edge connecting the two variables. If they are not, we will keep the edge. Based on the output from the equivalence class, the final step is to orient the other edges. It is important to check for conditional independence when removing edges and after, when changing the orientation of the arrows to make sure that they are not conditionally dependent, in order to check for colliders. It is important to keep in mind unmeasured variables that can be the cause of both X and Y to be dependent, even if we cannot find a Z.

In our study, an unmeasured confounder of the treatment-outcome and the mediator-outcome is the media awareness variable. Just to refresh, the media awareness variable represents the idea that the more access people have to social media, television, and news the more they will hear about COVID-19 which will influence how they act. Media awareness, U, has varying degrees of association with the treatment A, wearing masks, and outcome Y,

---

2 https://docs.google.com/presentation/d/1ckMGQ9YbjrUM5JCZ7EdNLrVfpS9pLYV0JUYMo9P0-xU/edit#slide=id.g704bd04db1_0_264

contracting COVID-19. In our study, we plan to measure media awareness quantitatively. This will allow us to find U using a linear regression model where the predictors are A and Y. The coefficients on the predictors determine the magnitude between U and A and U and Y. We do not have data on U so the plot from the sensitivity analysis allows us to see "how the ACE estimate changes as the magnitude between U and A and between U and Y changes."[3] The results from the sensitivity analysis and plot show how strong U would have to be to really change our conclusions.

Slide 9:

In step 6, we were instructed to create an outline for a quasi-experimental study design to estimate the effect of wearing masks to reduce the chances of contracting COVID-19. We thought that we could best apply an interrupted time series model where Y is the rate of contracting COVID-19. In an interrupted time series there is an intervention that occurs at a certain time. For example, a law that requires all civilians to wear a mask when entering essential businesses, that is put into effect in April 2020. Our main interest here is the effect of the mask law on the rate of COVID-19 cases. For this intervention, we hope that we can clearly see a noticeable change in Y and that the intervention is random enough "such that the trends in the outcome over time are comparable just before and after the intervention."[4]

We can see in the plot here that there is a discontinuity in the trend, therefore we may be able to assume that the intervention played a role. However, there are some limitations to an

[3] Myint, 09-sensitivity-solutions.html.
[4] Myint, 7:22 Quasi-Experimental Study Designs.

interrupted time series so we should not always assume that the law is the sole reason for any discontinuity. If our evidence is not strong enough that it was the law that caused this change instead of another factor then we should not claim that the enactment of a law to wear facemasks reduced the rate of contracting COVID-19. Perhaps, a change in the weather, from below freezing to low 70's caused the rate to drop. The severity or time of implementation of stay-at-home orders could be a confounding event that might be a reason that the number of cases of COVID-19 would drop. It would also be difficult to identify control states because not all states have the same exposure to the virus and have different policies regarding face masks and stay-at-home orders so it could be difficult to generalize for the United States.

Slide 10:

In order to answer the question, "Is it appropriate or useful to conduct this investigation in a time-varying treatment setting?" we need to first indicate a research question that could be addressed in this setting. The research question that we created is "Do transmission rates of COVID-19 affect the propensity that people wear masks in the future?" The treatment in this investigation is wearing masks. The time-varying confounder, L, is the transmission rate of COVID-19. The outcome, Y, is the chance one is infected by the virus.

In a time-varying treatment setting the treatment depends on some "bio-marker" and the treatment at a previous time point. We are curious about how the treatment changes as time passes. Treatment-confounder feedback might open up non-causal paths over time, so we want to ask "If the transmission rate is such a time-varying confounder variable, is it feasible to measure

that over time so that we might account for it?" KG and I agreed that the transmission rate is something that can be measured over time. Therefore, it is appropriate and useful to conduct this investigation in a time-varying treatment setting. We think that if the transmission rate of COVID-19 increases then, in the future, people will be more likely to wear masks. If the transmission rate decreases, people will be less likely to wear masks. At the same time, this change in the likelihood that people will or will not be more likely to wear masks will affect the future transmission rate of COVID-19.