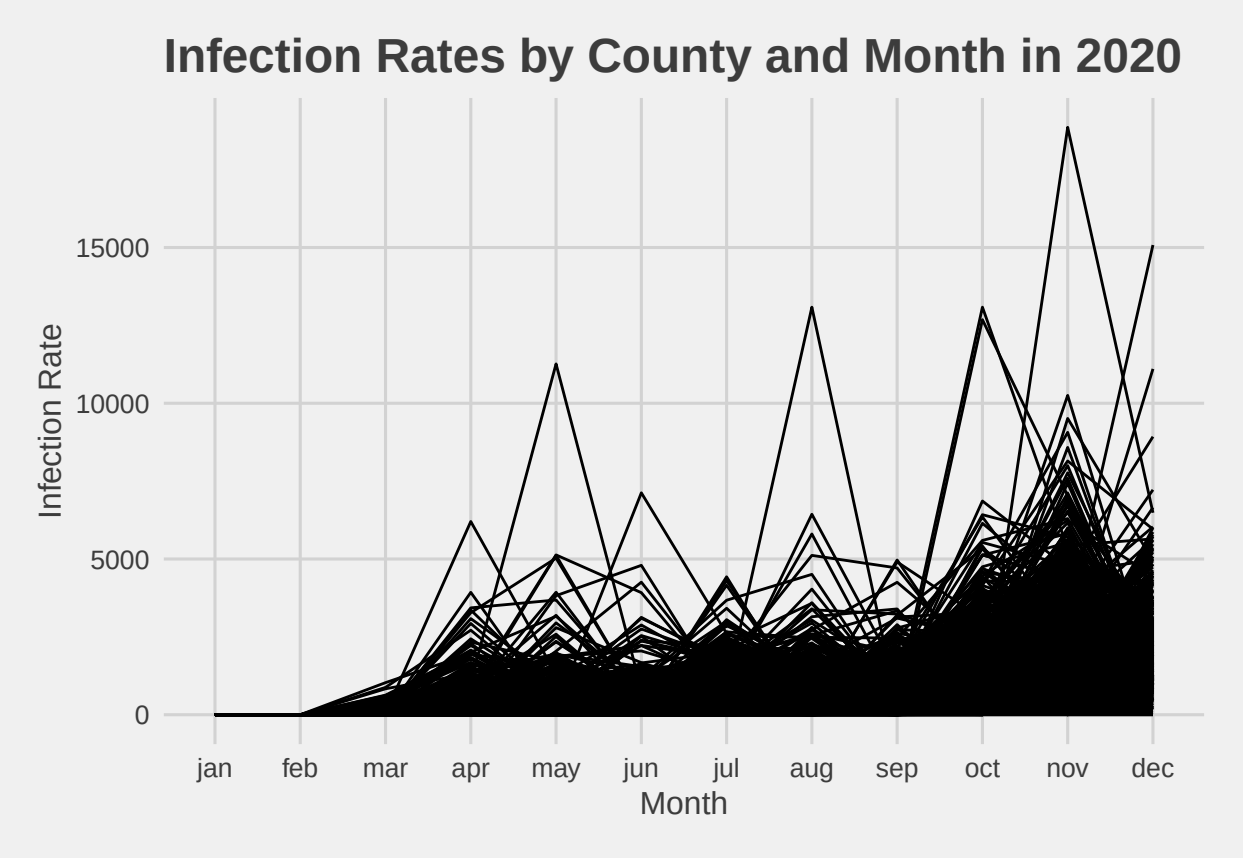
Introduction to Causal Inference - Final Project

Preprocessing Data

```
# transforming to a long dataset
infect <- covid_df %>%
  dplyr::select(countyFIPS, infection_jan, infection_feb, infection_mar, infection_apr, infection_may,
         infection_jun, infection_jul, infection_aug,infection_sep, infection_oct, infection_nov, infec

infect_long <- gather(infect, month, infection_rate, c(infection_jan, infection_feb, infection_mar, inf
                                                       infection_may, infection_jun, infection_jul, infection_a
                                                       infection_sep, infection_oct, infection_nov, infection_
                                                       ), factor_key=TRUE)
infect_long <- infect_long %>%
  arrange(countyFIPS)

# renaming months
infect_long %<>%
  separate(month, c(NA, "month"))

# joining with geographic and demographic data
wide_df <- covid_df[c(2:6,31:41)]
df <- left_join(x = infect_long,
                y = wide_df,
                by = "countyFIPS")

# rearranging our dataset
df <- df[c(1,4,5,6,2,3,7:18)]

# adding numerical month col
df %<>%
  mutate(mon = ifelse(month=="jan", 1, NA),
         mon = ifelse(month=="feb", 2, mon),
         mon = ifelse(month=="mar", 3, mon),
         mon = ifelse(month=="apr", 4, mon),
         mon = ifelse(month=="may", 5, mon),
         mon = ifelse(month=="jun", 6, mon),
         mon = ifelse(month=="jul", 7, mon),
         mon = ifelse(month=="aug", 8, mon),
         mon = ifelse(month=="sep", 9, mon),
         mon = ifelse(month=="oct", 10, mon),
         mon = ifelse(month=="nov", 11, mon),
         mon = ifelse(month=="dec", 12, mon))
```

```
# looking at infection rates over time
ggplot(df, aes(x=month, y=infection_rate, group = countyFIPS)) +
  geom_line() +
  theme_fivethirtyeight() +
```

```
  theme(axis.title = element_text()) +
  ggtitle('Infection Rates by County and Month in 2020') +
  xlab('Month') +
  scale_x_discrete(limits=c("jan","feb","mar","apr","may", "jun", "jul", "aug", "sep", "oct", "nov", "d
  ylab('Infection Rate')
```



In order to assess the impact of statewide mask mandates over time, we need to refine the dataset such that our treatment indicator shows the *month* each state passed the mandate (rather than just whether or not they passed it). We created a supplemental dataset based on news reports which identifies (1) the month that each state passed statewide mandates or (2) the month that the first city/county passed a mask mandate for those states with mandates in some parts.

```
# load dataset of when states passed their statewide mandates
states <- read_csv("data/state_mandates_month.csv")
```

```
## Parsed with column specification:
## cols(
##   state = col_character(),
##   mask_sum = col_double(),
##   date = col_character(),
##   source = col_character()
## )
```

```
states %<>%
  mutate(treat_date = ifelse(date=="apr", 4, NA),
         treat_date = ifelse(date=="may", 5, treat_date),
         treat_date = ifelse(date=="jun", 6, treat_date),
```

```
         treat_date = ifelse(date=="jul", 7, treat_date),
         treat_date = ifelse(date=="aug", 8, treat_date),
         treat_date = ifelse(date=="nov", 11, treat_date))
states_month <- df %>%
  group_by(state) %>%
  dplyr::select(state, month)

states_month <- left_join(states_month,states,by="state")
# deleting duplicates
states_month %<>%
    distinct(state, month, .keep_all = TRUE)

# adding numerical month col
states_month %<>%
  mutate(mon = ifelse(month=="jan", 1, NA),
         mon = ifelse(month=="feb", 2, mon),
         mon = ifelse(month=="mar", 3, mon),
         mon = ifelse(month=="apr", 4, mon),
         mon = ifelse(month=="may", 5, mon),
         mon = ifelse(month=="jun", 6, mon),
         mon = ifelse(month=="jul", 7, mon),
         mon = ifelse(month=="aug", 8, mon),
         mon = ifelse(month=="sep", 9, mon),
         mon = ifelse(month=="oct", 10, mon),
         mon = ifelse(month=="nov", 11, mon),
         mon = ifelse(month=="dec", 12, mon))

# creating treatment variable
states_month %<>%
  group_by(state) %>%
  mutate(treatment = ifelse(mon>=treat_date,1,0))

# assign SD as nontreated (the only state that never passed a mask mandate)
states_month %<>%
  mutate(treatment = ifelse(state=="SD",0,treatment))

# cleaning up dataframe
state_df <- states_month[c(1,2,6,8)]

# join with main df
df <- left_join(x = df,
                  y = state_df,
                  by = c("state", "month"))

# load kansas county dataset
kansas <- read_csv("data/kansas_counties.csv")
```

```
## Parsed with column specification:
## cols(
##   countyFIPS = col_double(),
##   county_name = col_character(),
##   mask_mandate = col_double()
## )
```

```
# getting county fips that passed mandate
kansas_nomask <- kansas %>% subset(is.na(mask_mandate)) %>%
  dplyr::select(countyFIPS)

kansas_nomask <- dplyr::pull(kansas_nomask, countyFIPS)

# updating kansas counties that did not pass mask mandates
df %<>%
  mutate(treatment = ifelse(countyFIPS %in% kansas_nomask, 0, treatment))

# updating Tennessee counties that never had mask mandate
tn_nomask <- c("47021", "47043", "47101", "47081")
df %<>%
  mutate(treatment = ifelse(countyFIPS %in% tn_nomask, 0, treatment))
```

Kansas is a unique case with mask mandate data available by county. On July 2, 2020, the governor of Kansas issued a state mandate, effective July 3, requiring masks or other face coverings in public spaces. As of August 11, 24 of Kansas's 105 counties did not opt out of the state mandate. 81 counties opted out of the state mandate, as permitted by state law, and did not adopt their own mask mandate. These 81 counties are regarded as untreated in our dataset. Source: https://www.cdc.gov/mmwr/volumes/69/wr/mm6947e2.htm

Four Tennessee counties that have never had a mask mandate: Cheatham, Dickson, Lewis and Hickman. These are regarded as untreated in the dataest. Source: https://fox17.com/news/local/here-are-the-tennessee-counties-where-masks-are-mandated-right-now-thanksgiving-covid-19-travel-nashville-williamson-davidson-wilson-rutherford-sumner-montgomery-henry-robertson-wayne-warren

Mississippi was the first state to lift state-wide mask mandate in October 2020. https://www.forbes.com/sites/nicholasreimann/2020/09/30/mississippi-becomes-first-state-to-lift-mask-mandate/?sh=16d8c2667f13

Finally, because unemployment changed so radically in 2020 and may be an important confounder, we bring in monthly unemployment data from the BLS. https://www.bls.gov/web/metro/laucntycur14.txt

```
state_unemp <- read_csv("data/county_employment_rates.csv")
```

```
## Parsed with column specification:
## cols(
##   `LAUS Area Code` = col_character(),
##   StateFIPS = col_double(),
##   CountyFIPS = col_double(),
##   `Area Title` = col_character(),
##   Period = col_character(),
##   `Civilian Labor Force` = col_number(),
##   Employed = col_number(),
##   Unemployed = col_number(),
##   Unemp_Rate = col_double()
## )
```

```
## Warning: 624 parsing failures.
##  row                  col expected actual                                 file
## 9580 Civilian Labor Force a number      - 'data/county_employment_rates.csv'
## 9580 Employed             a number      - 'data/county_employment_rates.csv'
## 9580 Unemployed           a number      - 'data/county_employment_rates.csv'
## 9580 Unemp_Rate           a double      - 'data/county_employment_rates.csv'
## 9581 Civilian Labor Force a number      - 'data/county_employment_rates.csv'
## .... ..................... ........ ...... .................................
## See problems(...) for more details.
```

```r
# deleting preliminary data
state_unemp %<>%
  subset(Period!="Feb-21(p)")
state_unemp_sm <- state_unemp[c(4,5,6,8,9)]

state_unemp_sm %<>%
  separate(Period, c("month", "year"))
state_unemp_sm$month <- tolower(state_unemp_sm$month)

state_unemp_sm %<>% rename(county_name = `Area Title`)
# joining

df <- left_join(x = df,
                  y = state_unemp_sm,
                  by = c("county_name", "month"))

# deleting duplicates
df %<>%
    distinct(countyFIPS, month, .keep_all = TRUE)

# rearranging and saving final df file
df <- df[c(1:5,19,21,18,6,7:17,23,24,25,20)]

# assessing missingness overall
sum(is.na(df)) # 792

# problem is SD treat_date is NA - fixing this
df %<>%
  mutate(treat_date=ifelse(state=="SD",0,treat_date))
```

Specify the Scientific Question

What is the effect of state mandates to wear a mask on Covid-19 infection rates in US counties during 2020?

This is a causal question, and not a statistical question, because we are not just interested in the observed association between mask enforcement by states and Covid-19 infection rate. We want to know what infection rates would have been if mask enforcement by states would have been generated differently, for example, by mandating all adults to wear a mask, or mandating no adults to wear a mask. The target population are US counties from January to December 2020.

Specify a Casual Model

Endogenous variables: Endogenous variables are factors that inform the scientific question we're asking, and of which we have a certain level of knowledge. Here, we're including all observable factors that are present in our date, but we could potentially include some unobservable factors as well, which we will address later.

Y = Infection rates at the county-level (for each month from Jan to Dec 2020) The exposure (state-wide mask mandate) has three possible levels: $\mathcal{A} = 0, 1, 2$ These levels correspond to mask enforcement at the state-level (no, yes, and some parts) No: 66, Yes: 1989, Some Parts: 1051

W represents the following covariates: W_1 = Median income at the county level W_2 = % of Whites at the county level W_3 = % of Hispanic at the county level W_4 = % of African Americans at the county level W_5 = % of Asian Americans at the county level W_6 = % of residents who finished college or associate education at the county level W_7 = % unemployed (by month) at the county level **I think we're not able to make any exclusion restrictions or independence assumptions, can you think of any?**

Independence assumptions: economic structure or policies developed at the county level can all be shared common causes of median income, race and educational attainment.

DRAW DAG

Exogenous variables: Unmeasured confounders: The exogenous variables include all the unmeasured factors that determine the values that the endogenous variables take. U is a placeholder for everything we do not know.

In this study, the exogenous nodes are $U = (U_{W1}, U_{W2}, U_{W3}, U_A, U_Y) \sim P_U$.

U_A refers to the unmeasured confounders related to mask enforcement by the state. These could be: political pressure from different constituencies to either pass or abstain from passing a mask mandate. U_{W1} refers to the unmeasured confounders related to median income at the county level. These could be: U_{W2} refers to the unmeasured confounders related to the percentage of White residents at the county level. These could be: U_Y refers to the unmeasured confounders related to mask enforcement by the state. These could be: genetic factors, medications, health care access, and travel behavior.

Structural Causal Model $W_1 = f_{W1}(U_{W1})$ $W_2 = f_{W2}(W_1, U_{W2})$ $W_3 = f_{W3}(W_1, W_2, U_{W3})$ $W_4 = f_{W4}(W_1, W_2, U_{W3}, U_{W4})$ $W_5 = f_{W5}(W_1, W_2, U_{W3}, U_{W4}, U_{W5})$ $W_6 = f_{W6}(W_1, W_2, U_{W3}, U_{W4}, U_{W5}, U_{W6})$ $A = f_A(W_1, W_2, W_3, U_A)$ $Y = f_Y(W_1, W_2, W_3, A, U_Y)$

Translate your question into a formal target casual parameter defined using counterfactuals

Should we do an MSM with three levels of treatment? The intervention of interest is forcing each individual to have a range of mask enforcement $a \in A$. We could define A={no, some parts, yes}.

The counterfactuals of interest are $(Y_a : a \in A)$, where A are the set of mask enforcement levels of interest. The counterfactual Y_a is the mortality/infection rate if, possibly contrary to fact,the individual lived in a state with mask enforcement A=a from the January to December 2020 period.

We could also do yes versus any other level of treatment, and do the causal risk difference of treatment being yes versus everything else?

Specify your observed data and its link to the casual model

- Describe the data you are working with and its link to the casual model you have specified.

We assume the observed data were generating by sampling n times from a data generating system contained in the structural causal model MF, resulting in ni.i.d copies of the random variable O. This provides a link between the causal model MF and the statistical model M.

The statistical model M is the set of possible observed data distributions. We have not placed any restrictions on the statistical model, which is thereby non-parametric.

- Be sure to include a basic descriptive table of your data that provides information on the outcome, exposure, and covariate distributions.

Is this table more than this?

```
##
## Please cite as:

##  Hlavac, Marek (2018). stargazer: Well-Formatted Regression and Summary Statistics Tables.

##  R package version 5.2.2. https://CRAN.R-project.org/package=stargazer
```

Table 1: Descriptive Statistics

Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
Infection Rate	37,272	526.50	845.70	0	7.9	686.3	18,858
Population	37,272	102,729.20	332,782.90	169	10,867	67,029	10,039,107
Median Household Income	37,272	52,665.05	13,758.93	25,385	43,650	58,738	140,382
Population Density	37,272	240.97	1,723.23	0.04	16.38	113.94	71,890.61
Percent African American	37,272	9.24	14.35	0.00	0.89	10.63	86.59
Percent Latino	37,272	9.76	13.88	0.65	2.45	10.11	96.35
Percent Asian American	37,272	1.52	2.79	0.00	0.48	1.42	43.36
Percent White	37,272	84.60	16.18	3.76	79.86	95.44	99.04
Percent Native American	37,272	2.37	7.76	0.00	0.40	1.35	92.41
Percent Pacific Islander	37,272	0.12	0.39	0.00	0.03	0.12	12.84
Percent College	37,272	5.28	0.99	1.49	4.61	5.93	9.30
Unemployment Rate	37,272	6.75	3.79	0.70	4.10	8.30	39.70

Identify

Is your target casual parameter identified under your initial causal model? If we're not making any assumptions on independence or exclusion restrictions, then we won't be able to identify the causal parameter under our initial causal model.

If not, under what additional assumptions would it be identified? How plausible are these for your particular problem? Are there additional data/changes to your study design that would improve their plausibility?

Commit to a Statistical Model and Estimand (Target parameter ofthe observed data distribution).

State these explicitly.

Estimate.

Apply each of the three estimators we have learned in class (simple or non-targeted substitution estimator-a.k.a G-computation estimator), Inverseprobability of treatment weighted estimator, and TMLE) to estimate yourtarget parameter. Use of thetmle,ltmle, or otherRpackages is accept-able. Also report unadjusted results for comparison.

Use Super Learner when implementing TMLE. For comparison, you maywish to use it when implementing your G-computation and IPTW esti-mators also. A simple library is fine (writing wrappers to include yourown parametric regressions as candidates is great). Include an assessmentof the performance (cross validated risk) of the algorithms in your library.It is helpful to include the simple mean as a benchmark. Also report anestimate of the cross-validated risk of the SL and interpret.

Provide some formally assessment of the positivity assumption. Evalu-ate the distribution of your estimated propensity scoregn(A= 1|Wi), i=1, . . . , n, and corresponding non-stabilized weights (as well as of your sta-bilized weights if you use stabilized weights to fit an MSM). Considerevaluating sensitivity to different truncation levels forgn. Note that forTMLE, bounding or truncatinggnaway from 0 is recommended on thebasis of both theory and finite sample performance; for IPTW it can helpor hurt. Report how for how many observations wasgn(A|Wi) truncated.

Present a detailed plan for statistical inference/variance estimation basedon the non-parametric bootstrap, and implement it (understanding thattime may be a limitation depending on your SL library). When boot-strapping an estimator that uses cross-validation to estimate nuisance pa-rameters, be careful to ensure that all copies of a given independent unitare contained within the same fold. Plot your bootstrap distribution

and comment as appropriate. For TMLE (and IPTW), you can also report ainfluence curve based variance estimate for comparison, if you wish.

Interpret results.

What is the statistical interpretation of your analyses? Discuss differences (or lack thereof) in the estimates provided by the differentestimators. What is the causal interpretation of your results and how plausibleis it? What are key limitations of your analysis? How might these results (ifat all) inform policy, understanding, and/or the design of future studies?