Study of the influence of public policy decisions on Covid 19 infections

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

Section: 203.7

Professor: **Mark Labovitz**

Day & Time: Saturday 10 AM

Team Members (Team 3):

George Jiang

Ramon Jimenez

Sudhrity Mondal

Table of Contents

1.0 Introduction

The spread of Covid-19 across the globe has created a pandemic unseen in the past century. Governments, corporations, households and people around the world have made considerable changes in the way we operate and proceed with our day-to-day life. These changes in the US are governed by local and state level policies and vary from state to state.

The US federal government did not dictate any specific policy as a result the policy variations across different states were based on state leaders' view of the crisis, their beliefs and dependence on conclusions based on science etc.

Our causal theory is that various state policies have helped reduce the infection rate of Covid-19. Therefore, this research study focuses on the following question:

What is the effect size of state level public health policy decisions on the rate of COVID-19 infection after taking into account the variations across different states like population density, poverty rate and population under risk of serious illnesses?

1.1 Variables and Measurements

To reduce type 1 error inflation, we preselected the **type of causal variables** we are going to use in the dataset before we did any EDA. Any EDA in the dataset is used to determine the best variable to include for each type and which variable transformation to use in the final model. The type of causal variables we decided to include in our causal model: state policy, age, health, socioeconomics, homeless, population density, baseline infection rate before state policy, and personal beliefs about state policy.

Dependent variable

1. **Cases in Last 7 Days**: We have chosen *CasesinLast7Days* as the dependent variable to study the immediate effect of public policy decisions over a short period of time, because the cases in the last 7 days happened after the State policy took effect.

   A new variable **infection_last_7_rate**=*CasesinLast7Days*/*Population2018**10000 is created and used in the analysis, because we want to control for the difference in population across different States. Furthermore, we multiply the infection rate by 10000 to convert the percentage to basis point, because it makes interpretation of the coefficients easier.

   Variable Type: Real Number

Independent variables related to causal theory where we have data

1. **Stay at home/ shelter in place**: The variables *Stayathomeshelterinplace* and *Endstayathomeshelterinplace* are used to calculate the duration of Stay at home/shelter in place in days. This is then converted to a binary variable. This is the key explanatory variable we want to estimate the effect size. A new variable **stay_at_home** is created and used in the analysis. Note, because 3 States had *Endstayathomeshelterinplace* date but no *Stayathomeshelterinplace*, we set these 3 samples as NA, because we don't know whether these 3 States had a stay_at_home policy.

   Variable Type: Binary

2. **Mandate face mask use by all individuals in public spaces**: The variables *Mandatefacemaskusebyallindividualsinpublicspaces* and *Stateendedstatewidemaskusebyindividualsinpublicspaces* are used to calculate the duration of mandated face masks in days. This is then converted to a binary variable. This policy variable serves as a control to better estimate the effect size of **stay_at_home**. A new variable **mask_mandate** is created and used in the analysis.

   Variable Type: Binary

3. Population density per square miles: The population density (*Populationdensitypersquaremiles*) is used as a control variable because population density and Covid-19 infection are closely related. A new variable **pop_density**=Populationdensitypersquaremiles is created and used in the analysis.

   Variable Type: Real Number

4. Percent at risk for serious illness due to COVID: The percent at risk for serious illness due to Covid-19 (*PercentatriskforseriousillnessduetoCOVID*) is used as control because high risk population affects both the **stay_at_home** policy and infection rate. A new variable **due_covid_serious_ill_rate**=*PercentatriskforseriousillnessduetoCOVID* is created and used in the analysis.

   Variable Type: Real Number (0-100)

5. Number Homeless (2019): The variable *NumberHomeless2019* and *Population2018* are used as a control variable because they do not have a shelter in place. A new variable **homeless_2019_rate**=*NumberHomeless2019*/*Population2018**10000 is created and used in the analysis.

   Variable Type: Real Number

6. **Median Annual Household Income**: The variable *MedianAnnualHouseholdIncome* is used as a control for Covid-19 infection as household income level dictated if people had the means to support themselves and to adhere to the shelter in place policy guidelines. A new variable **median_annual_household_income**=*MedianAnnualHouseholdIncome* is created and used in the analysis.

   Variable Type: Real Number

1.2 Potential Issues & Considerations

- There are two omitted variables (baseline infection rate and personal beliefs) that we cannot find in the dataset, which will bias our beta in our model.

- Data is not identically distributed with the current state metrics: State population, population density and homeless numbers may have changed considerably since 2018 or 2019 and the data isn't updated for 2020.

- Sample size is only 51, therefore we might not have enough statistical power to detect large enough differences in state policy.

2.0 Exploratory Data Analysis

2.1 Read and Review Data

The data file used for the research is **covid-19.xlsx**. Relevant data is extracted from the data for analysis.

```
# Read excel data

#https://drive.google.com/file/d/1MDRtdf-UI50lxWTJc4XMiif0NuanHef9/view?usp=sharing
system("gdown --id 1MDRtdf-UI50lxWTJc4XMiif0NuanHef9")
full_data=read_excel('covid-19.xlsx', sheet=2, skip=1, col_names=TRUE)
```

2.2 Column Renames and Data Processing

The input data is processed to create ratios where appropriate. The column names are changed to a name that is more representative of the data content.

```
# Renaming column names and feature extraction
data<-full_data%>%
  transmute(
    #rates
    state=State,
    infection_last_7_rate=CasesinLast7Days/Population2018*10000,
    infection_rate=TotalCases/Population2018*100,
    covid_death_last_7_rate=DeathsinLast7Days/CasesinLast7Days*10000,
    covid_death_rate=TotalDeaths/TotalCases*100,
    test_rate=XTestsPerformed/Population2018*100,
    white_pop_rate=WhiteofTotalPopulation14*100,
    black_pop_rate=as.numeric(BlackofTotalPopulation16)*100,
    hispanic_pop_rate=as.numeric(HispanicofTotalPopulation18)*100,
    other_pop_rate=OtherofTotalPopulation28*100,
    due_covid_serious_ill_rate=PercentatriskforseriousillnessduetoCOVID,
    non_elderly_pec_rate=NonelderlyAdultsWhoHaveAPreExistingCondition/Population2018*100,
    all_death_2018_rate=Allcausedeaths2018/Population2018*100,
    unemployed_2018_rate=PercentUnemployed2018,
    poverty_rate=Percentlivingunderthefederalpovertyline2018,
    homeless_2019_rate=(NumberHomeless2019/Population2018)*10000,
```

```
    pop_under18_rate=Children018*100,
    pop_19_25_rate=Adults1925*100,
    pop_26_34_rate=Adults2634*100,
    pop_35_54_rate=Adults3554*100,
    pop_54_64_rate=Adults5564*100,
    pop_65_plus_rate=X65*100,
    Retailrecreation_rate=Retailrecreation,
    Grocerypharmacy_rate=Grocerypharmacy,
    Parks_rate=Parks,
    Transitstations_rate=Transitstations,
    Workplaces_rate=Workplaces,
    Residential_rate=Residential,
    medicaid_exp_rate=as.numeric(MedicaidExpendituresasaPercentofTotalStateExpendituresbyFund)*100,
    case_rate=CaseRateper100000,
    case_rate_last_7=CaseRateper100000inLast7Days,

    #absolute numbers
    homeless_2019=NumberHomeless2019,
    pop_density=Populationdensitypersquaremiles,
    poverty=Percentlivingunderthefederalpovertyline2018/100*Population2018,
    pop_under18=Children018*Population2018,
    pop_19_25=Adults1925*Population2018,
    pop_26_34=Adults2634*Population2018,
    pop_35_54=Adults3554*Population2018,
    pop_54_64=Adults5564*Population2018,
    pop_65_plus=X65*Population2018,
    median_annual_household_income=MedianAnnualHouseholdIncome,
    federal_aid=WeeklyUImaximumamountwithextrastimulusthroughJuly312020dollars,
```

```
    #policy dates

    business_close_dt=as.Date(Closedothernonessentialbusinesses,origin = "1899-12-30"),
    business_reopen_dt=as.Date(Begantoreopenbusinessesstatewide,origin = "1899-12-30"),
    stay_at_home_dt=as.Date(Stayathomeshelterinplace,origin = "1899-12-30"),
    end_stay_at_home_dt=as.Date(Endstayathomeshelterinplace,origin = "1899-12-30"),
    mask_mandate_dt=as.Date(Mandatefacemaskusebyallindividualsinpublicspaces,
        origin = "1899-12-30"),
    end_mask_mandate_dt=as.Date(Stateendedstatewidemaskusebyindividualsinpublicspaces,
        origin = "1899-12-30"),
        stay_at_home_dt=as.Date(ifelse(stay_at_home_dt=='1899-12-30','2020-10-30',
        as.character(stay_at_home_dt))),
    end_stay_at_home_dt=as.Date(ifelse(end_stay_at_home_dt=='1899-12-30','2020-10-30',
        as.character(end_stay_at_home_dt))),
    mask_mandate_dt=as.Date(ifelse(mask_mandate_dt=='1899-12-30','2020-10-30',
        as.character(mask_mandate_dt))),
    end_mask_mandate_dt=as.Date(ifelse(end_mask_mandate_dt=='1899-12-30','2020-10-30',
        as.character(end_mask_mandate_dt))),
```

```
    #policy duration
    stay_at_home_duration=as.numeric(end_stay_at_home_dt-stay_at_home_dt),
    mask_mandate_duration=as.numeric(end_mask_mandate_dt-mask_mandate_dt),
    business_close_duration=as.numeric(business_reopen_dt-business_close_dt),

    #policy enactment binary
    legal_enforcement_mask=ifelse(mask_mandate_duration>0 &
        Nolegalenforcementoffacemaskmandate==0,1,0),
    business_close=ifelse(business_close_duration>0,1,0),
    stay_at_home=ifelse(stay_at_home_duration>0,1,0),
    mask_mandate=ifelse(mask_mandate_duration>0,1,0)

  )

data$stay_at_home <- replace(data$stay_at_home, which(data$stay_at_home_duration <0), NA)

data$stay_at_home_duration <- replace(data$stay_at_home_duration,
    which(data$stay_at_home_duration < 0), NA)
```

We processed data for the following reasons:

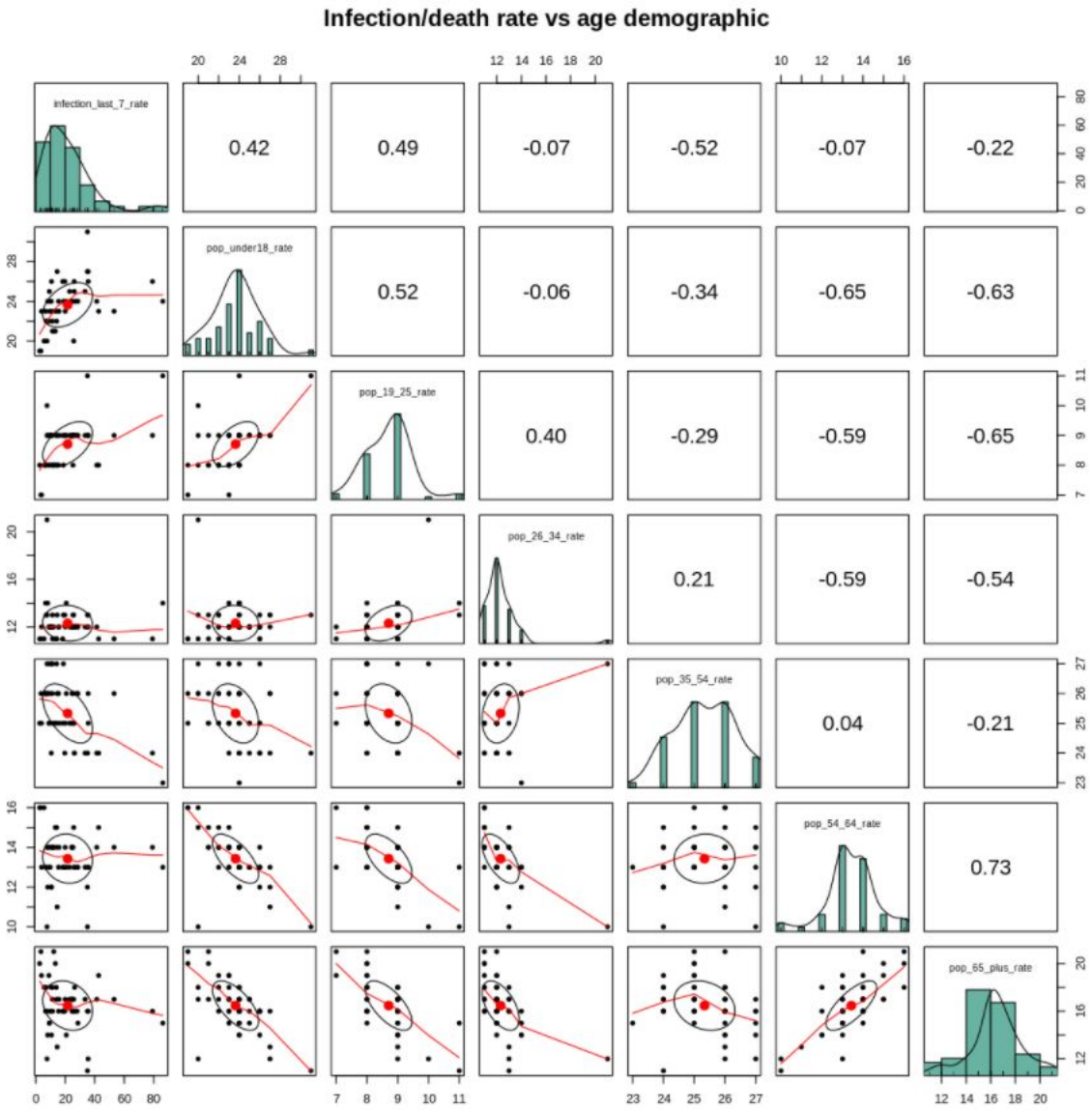
- We renamed long names to shorter names that is easier to understand
- We converted most data to percentage to control for difference in population between States
- We multiplied these percentages by 100 to make the coefficients larger easier to interpret.
- We checked for incorrect or weird values and found 3 States with missing stay at home policy data, in the end we removed them. But in hindsight, we should have kept them and imputed the correct data to avoid deletion bias.
- We didn't remove any "outliers", because we don't want to introduce any subjective bias into the data.

2.3 Pairs Review for Feature Selection

After data processing, we use pairs function to do exploratory data analysis (EDA). By using pairs, we can examine each variable's relationship with the target outcome variable, including shape of the variable distribution, scatterplot and correlation, which we will use for feature selection later on. Furthermore, we use pairs on both the original data and transformed data to compare the which variables we will include in our final model.

Infection rate vs age demographic

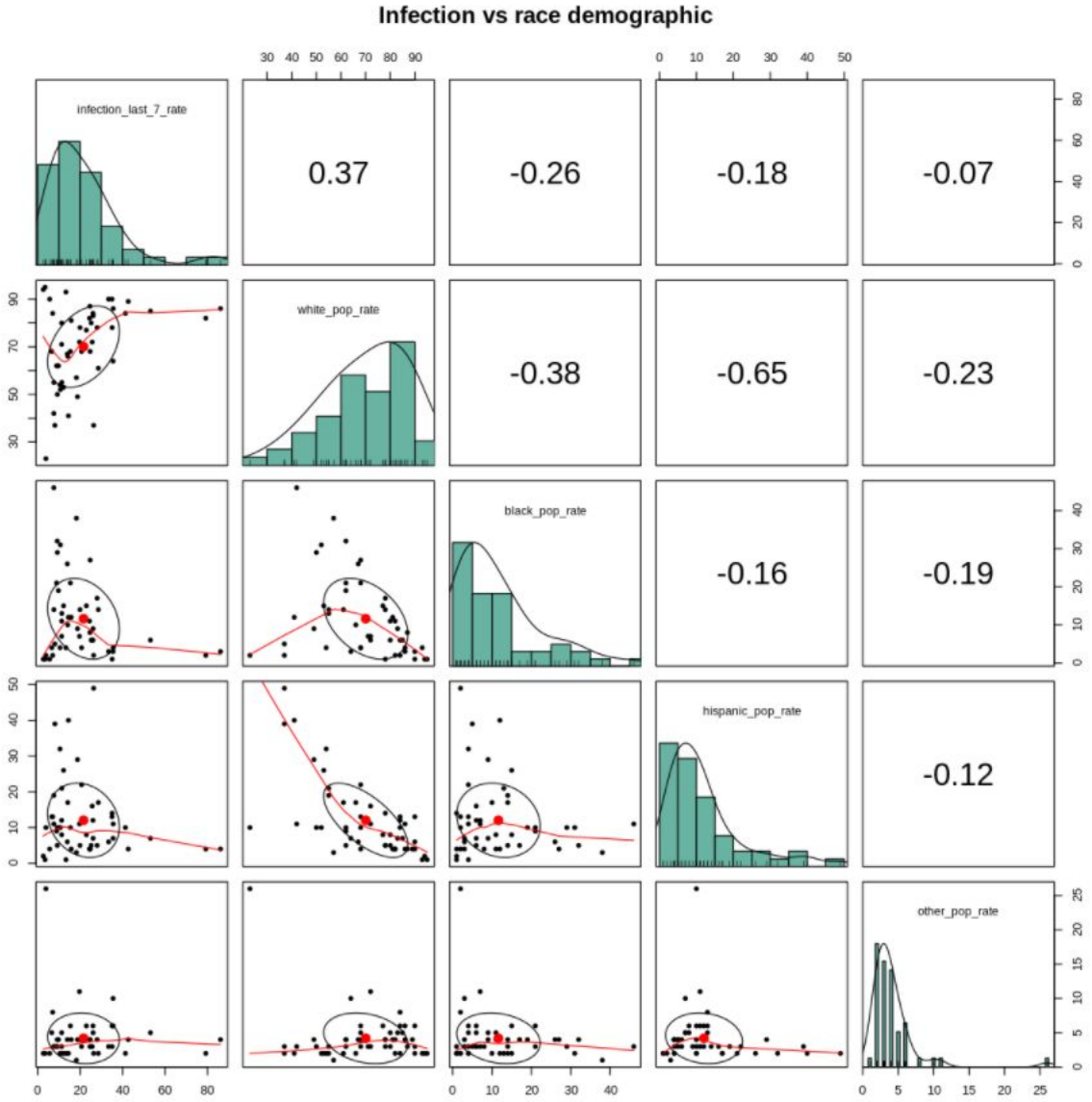
```
#Pairs of relevant feaures

#infection/death rate vs age demographic
age<-data%>%
  select(infection_last_7_rate,pop_under18_rate,pop_19_25_rate,
    pop_26_34_rate,pop_35_54_rate,pop_54_64_rate, pop_65_plus_rate)

pairs.panels(age, method = "pearson", hist.col = "#69b3a2",
main = "Infection/death rate vs age demographic")
```



Comments – Age factor:

- Age variables are approximately normally distributed.
- Age group <= 25 and 35-54 show moderate/strong correlation with last 7 days infection rate. This makes logical sense as we can reasonably argue that younger populations are more likely to go out and get infected, therefore, we might control for their population difference across the States.
- In addition, these group's scatterplot shows a linear relationship with infection rate
- We will consider age group for our control variable

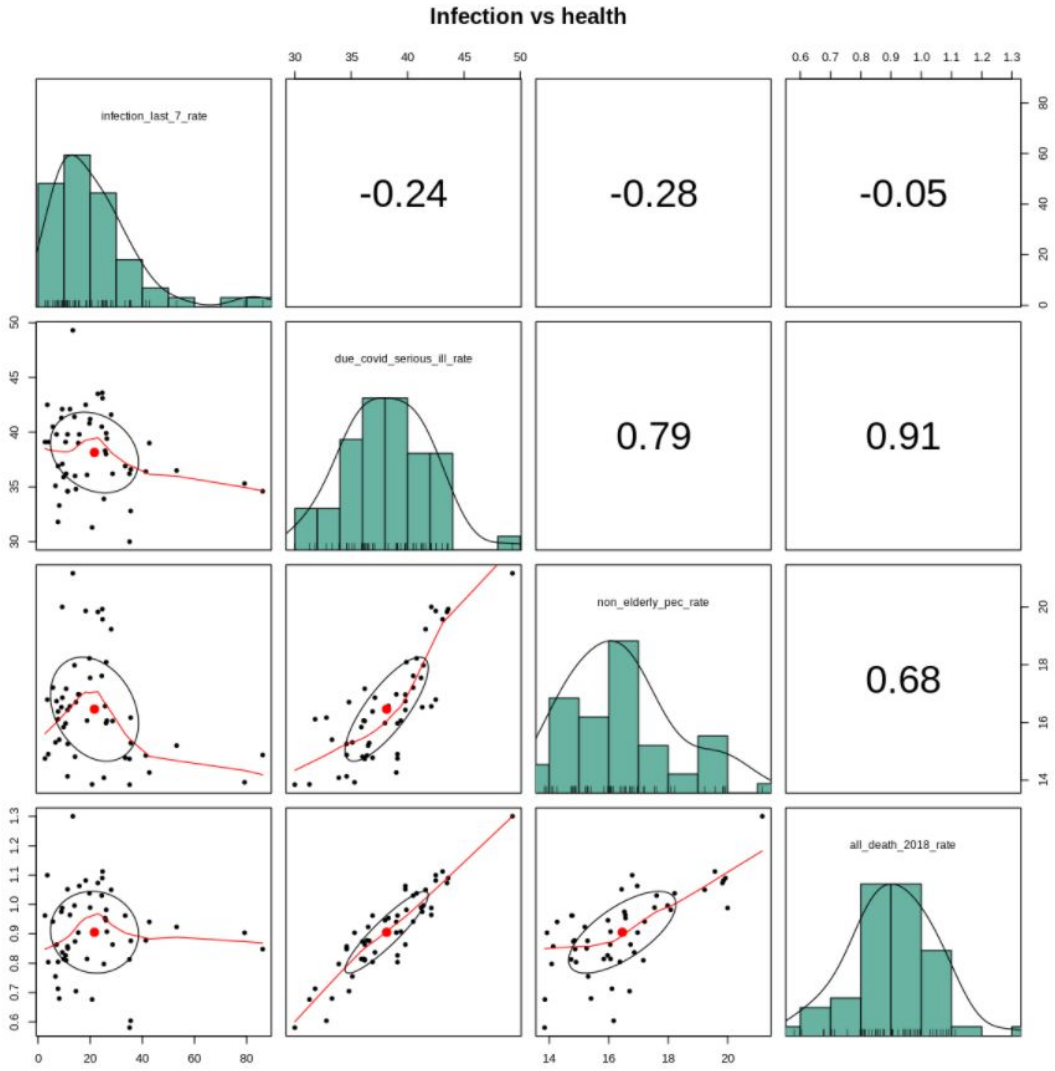
Infection rate vs age demographic



Comments – Race factor:

- White population rate shows moderate correlation with last 7 days infection rate
- However, the race group's scatter plot doesn't show a linear relationship with infection rate even after log transformation, which violates the assumption of CLM.
- Lastly, we don't see any causal link between race and infection that is not already explained by other socioeconomic factors, therefore we will not use Race as a controlling variable, but we will include in model 3 to overfit the model per lab 2 instruction.

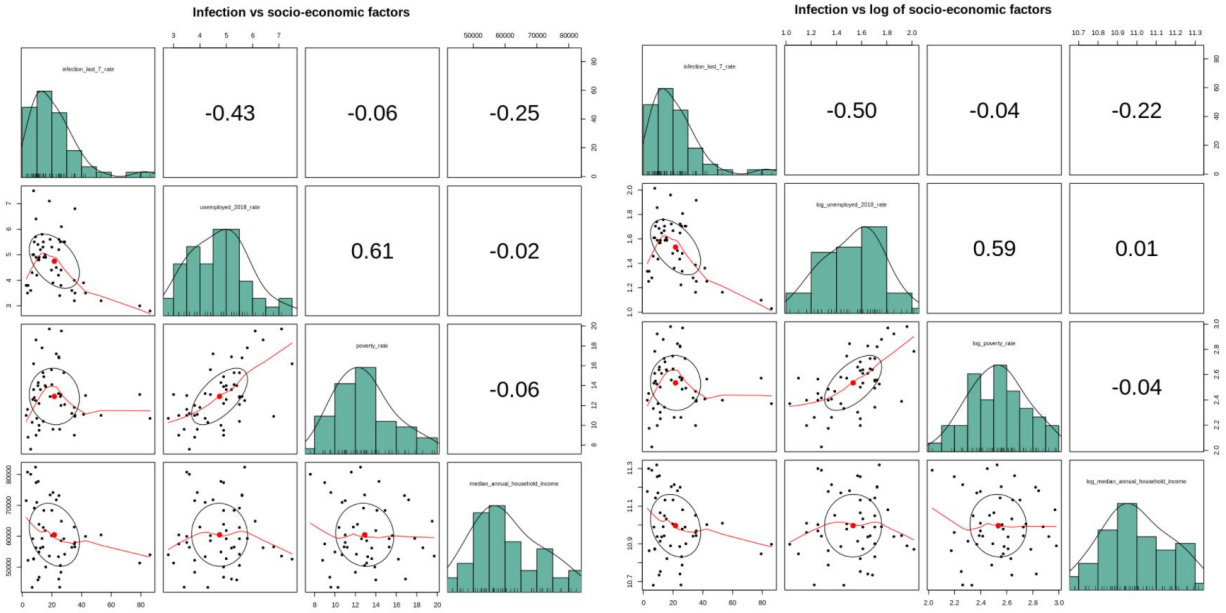
Infection rate vs Health



Comments – Health factor:

- Health variables are approximately normally distributed.
- Both due_covid_serious_ill_rate and non_elderly_pec_rate show moderate correlation with last 7 days infection rate.
- However, only due_covid_serious_ill_rate's scatterplot shows a more linear relationship with infection rate.
- In addition, due_covid_serious_ill_rate represents a % population that is at high risk to covid-19, who are more likely to self-quarantine even without **stay_at_home** policy. Therefore, we should use it as a control variable.
- Because this variable already includes age information, we consider using it over age variables because it can control for 2 variables with a single parameter.

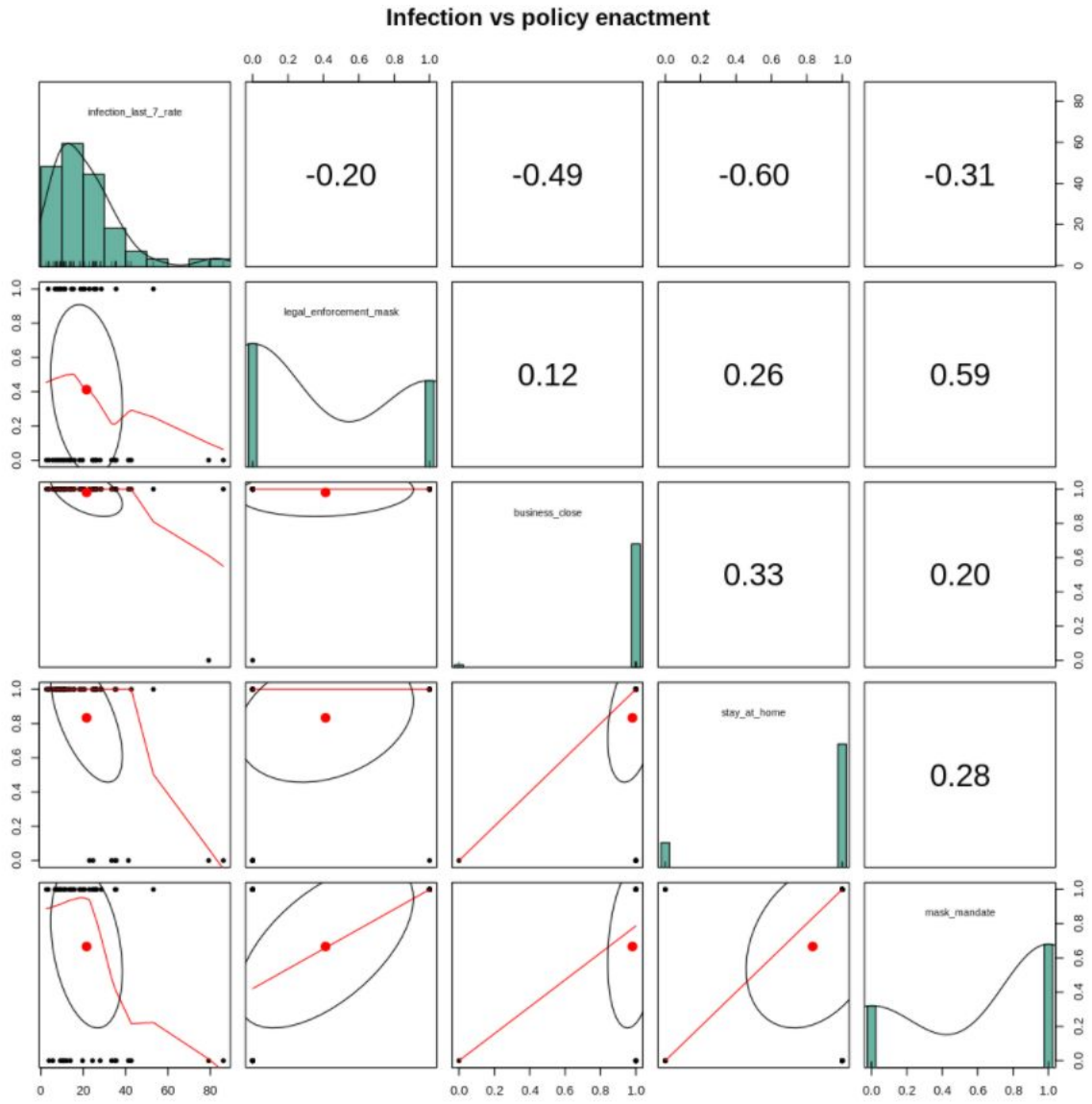
Infection rate vs socio-economic factors



Comments – Socio-economic factors (unemployed_2018, poverty_rate, median_household income):

- Socio-Economic variables are approximately normally distributed
- Both unemployment rate and median household income show moderate correlation with the last 7 days infection rate.
- However, only median household income shows some linear relationship between it and outcome variables.
- Furthermore, unemployment rate is based on 2018 data, which may have changed in 2020.
- In addition, income does play an effect on infection rate as lower income families cannot afford to self-quarantine. Therefore, we consider using median household income as a control variable.
- We notice log(poverty_rate) shows a much stronger correlation with infection than poverty rate, therefore, we decide to use log(poverty_rate).

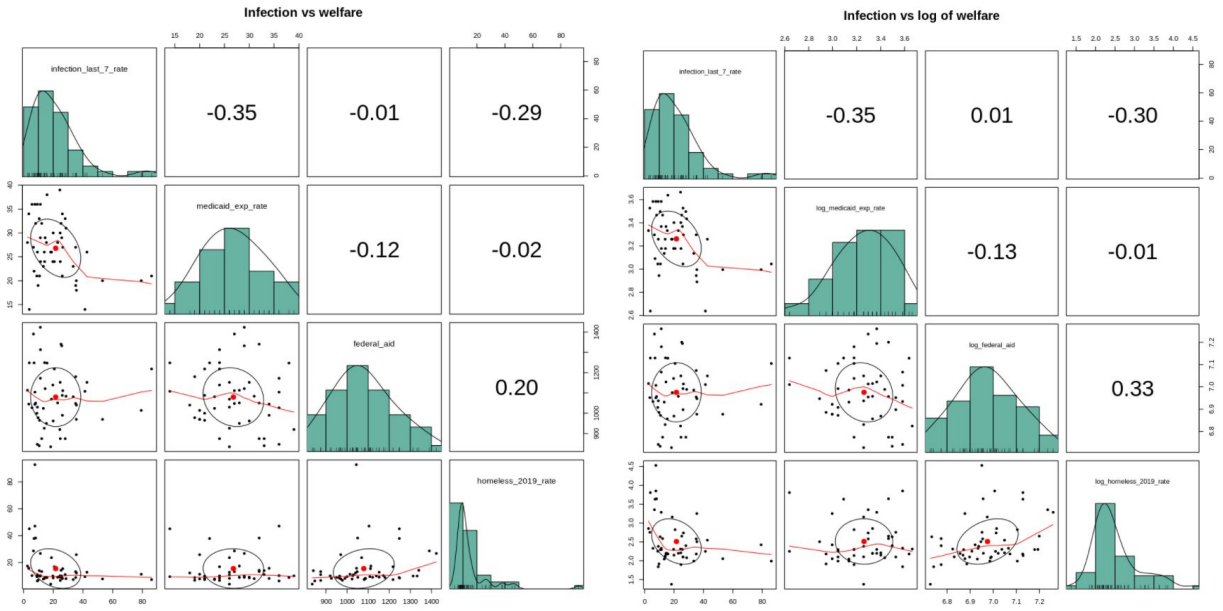
Infection/death rate vs State Policies



Comments – State Policies enactment (enforcement, business_close, stay at home, mask):

- Binary variables by nature have a linear relationship with outcome variables.
- Although business close show strong correlation with infection rate, because 50 out of the 51 States enacted business close, we don't have enough variance and degree of freedom to estimate the effect size of business close
- legal enforcement and mask mandate are highly related variables, we have to pick one, we picked mask mandate because it shows a stronger correlation.
- Therefore, we will only estimate the effect size of stay_at_home, and mask mandate

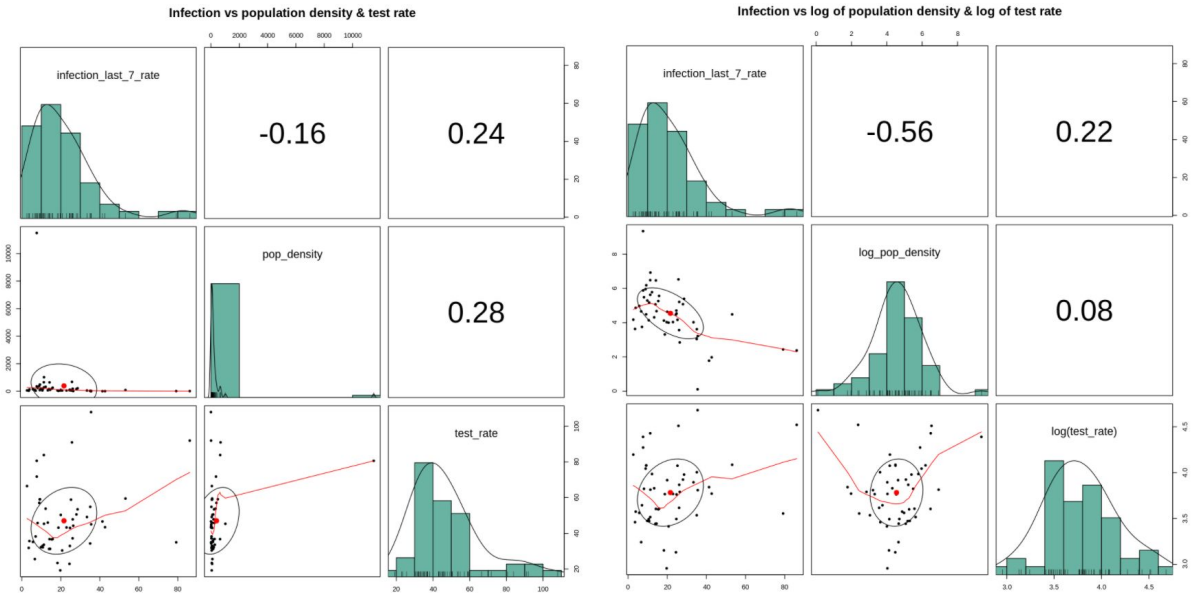
Infection rate vs Welfare



Comments – Welfare (medicaid, federal_aid, homeless_2019):

- Both Medicare expenditure rate and homeless rate show moderate correlation with infection rate
- However, Medicare expenditure rate doesn't seem to have a causal link to infection rate, while homeless rate does to a causal link as they are one of the more vulnerable populations
- After log transformation, the homeless rate appears to have a more linear relationship with infection rate.
- Therefore, we will consider it for our control variable.

Infection rate vs Welfare



Comments – Population density and test rate:

- After log transformation pop density shows a strong correlation with infection rate.
- It also has a causal link with the infection rate
- Furthermore, it shows a linear relationship with the outcome variable
- Therefore, we will include it in our control variable.

2.4 Feature extraction & Review of features

The X and Y variables are extracted to get summary information.

```
# Dependent and Independent Variables

y = c("infection_last_7_rate")
X = c("stay_at_home", "mask_mandate", "pop_density", "due_covid_serious_ill_rate",
    "homeless_2019_rate", "median_annual_household_income")
oth = c("state")
# Extract relevant variables from partially processed data
data_set_core <- data[,c(oth, y, X)]
summary(data_set_core)
head(data_set_core)
```

```
   state           infection_last_7_rate  stay_at_home     mask_mandate
Length:51          Min.   : 2.682        Min.   :0.0000   Min.   :0.0000
Class :character   1st Qu.:10.185        1st Qu.:1.0000   1st Qu.:0.0000
Mode  :character   Median :18.282        Median :1.0000   Median :1.0000
                   Mean   :21.621        Mean   :0.8333   Mean   :0.6667
                   3rd Qu.:26.252        3rd Qu.:1.0000   3rd Qu.:1.0000
                   Max.   :86.097        Max.   :1.0000   Max.   :1.0000
                                         NA's   :3
 pop_density       due_covid_serious_ill_rate homeless_2019_rate
Min.   :    1.11   Min.   :30.00              Min.   : 3.964
1st Qu.:   48.66   1st Qu.:35.95              1st Qu.: 8.534
Median :   93.24   Median :38.30              Median :10.002
Mean   :  392.64   Mean   :38.15              Mean   :15.517
3rd Qu.:  209.56   3rd Qu.:40.65              3rd Qu.:14.842
Max.   :11496.81   Max.   :49.30              Max.   :92.832

median_annual_household_income
Min.   :43469
1st Qu.:53434
Median :58882
Mean   :60478
3rd Qu.:68380
Max.   :82372
NA's   :1
```

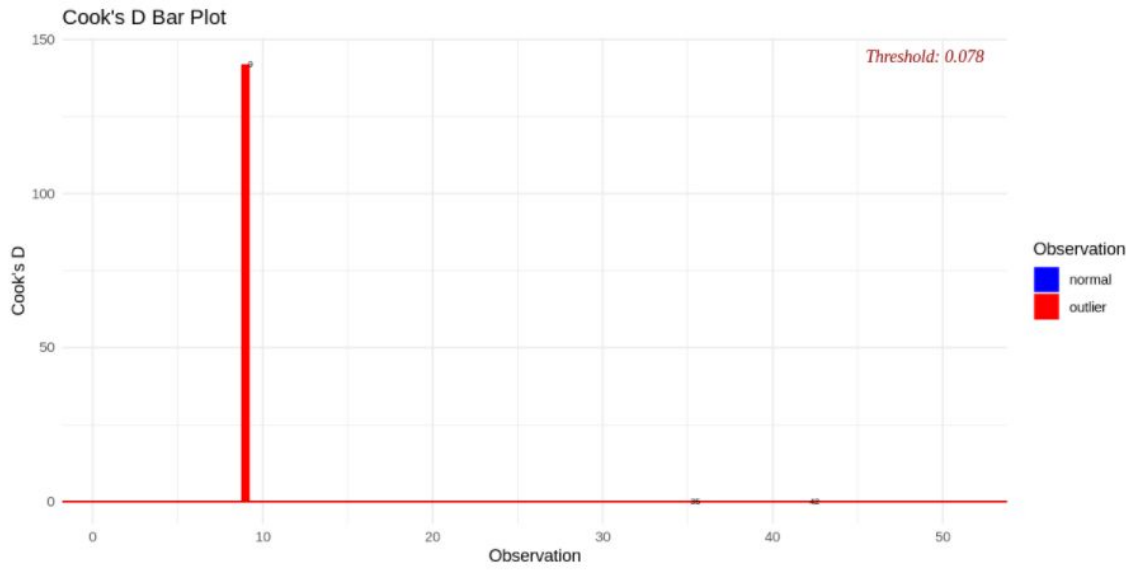
state	infection_last_7_rate	stay_at_home	mask_mandate	pop_density	due_covid_serious_ill_rate	homeless_2019_rate	median_annual_household_income
<chr>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
Alabama	24.724466	1	1	93.24	43.1	6.671616	73181
Alaska	35.514850	1	1	1.11	32.8	25.859801	56581
Arizona	10.561034	1	0	62.91	39.1	13.953561	45869
Arkansas	22.940947	0	1	56.67	43.5	9.015122	71805
California	8.135087	1	1	241.65	33.3	38.242998	69117
Colorado	20.703832	1	1	54.72	31.3	16.888582	74168

The data for stay_at_home, median_annual_household_income contains NA values, which will not be used during model creation.

Check for Measures of Influence

```
# Check for Measures of Influence

model <- lm(infection_last_7_rate~pop_density, data=data)
cutoff <- 4/((nrow(data) - length(model$coefficients)-2))

writeLines("Cook's distance plot for Measures of Influence\n")

options(repr.plot.height=5, repr.plot.width=10)
par(mfrow=c(1,2), mar=c(2,2,5,5), mgp=c(.8,.1,0))

ols_plot_cooksd_bar(model)
```

```
Cook's distance plot for Measures of Influence
```




```
a <- data_set_core %>%
  mutate(row_num = row_number())
a[c(9), 1:8]

data <-data[!(data$state == 'District of Columbia'),]
```

A tibble: 1 × 8

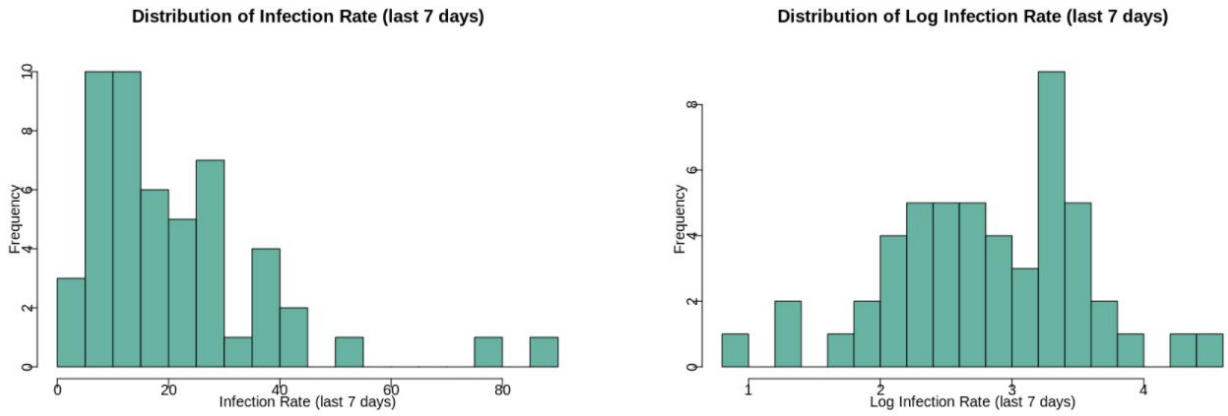
state	infection_last_7_rate	stay_at_home	mask_mandate	pop_density	due_covid_serious_ill_rate	homeless_2019_rate	median_annual_household_income
<chr>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
District of Columbia	7.644618	1	1	11496.81	31.8	92.83157	52594

We see that the District of Columbia has a lot of influence on the infection rate because of its population density. District of Columbia population density seems to be an outlier. However, we don't want to remove data just because it is skewed, therefore, we ended up log transforming the population density.

Histograms

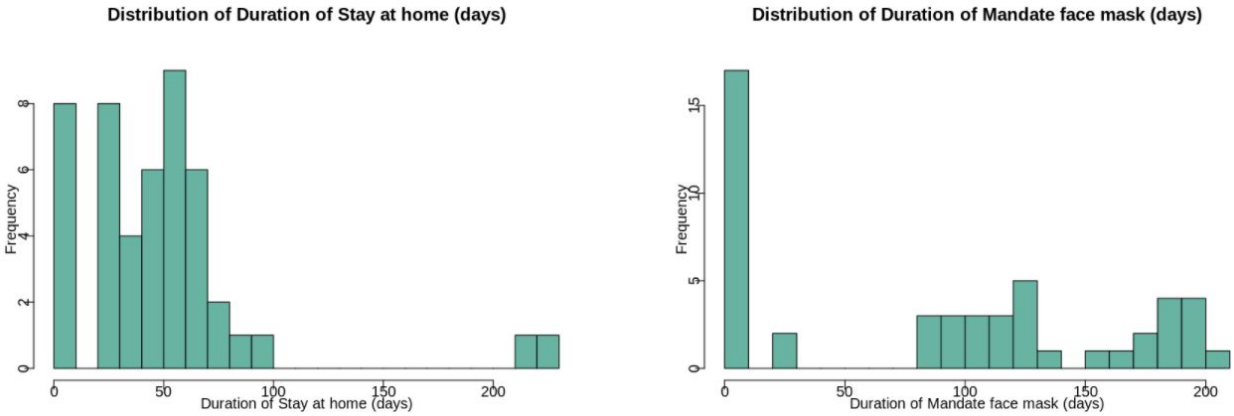
Infection Rate last 7 days

```
options(repr.plot.height=5, repr.plot.width=15)
par(mfrow=c(1,2), mar=c(2,2,5,5), mgp=c(.8,.1,0))
# Distribution of Dependent Variable - Infection Rate (last 7 days)
hist(data_set_core$infection_last_7_rate, main="Distribution of Infection Rate (last 7 days)",
    xlab="Infection Rate (last 7 days)", breaks=20, col="#69b3a2")
hist(log(data_set_core$infection_last_7_rate), main="Distribution of Log Infection Rate (last 7 days)",
    xlab="Log Infection Rate (last 7 days)", breaks=20, col="#69b3a2")
```



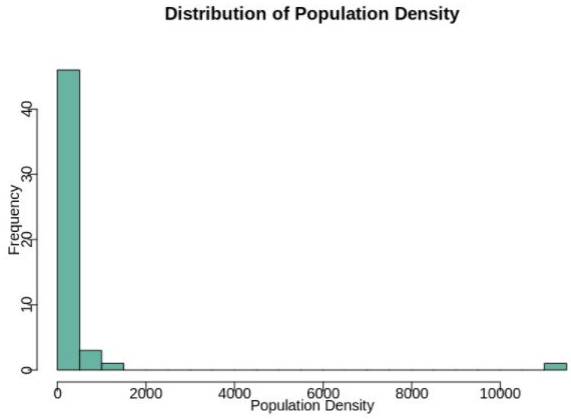
Comments: The distribution infection rate last 7 days seems to be multi-modal and skewed to the right. A log of the distribution makes it closer to normal.

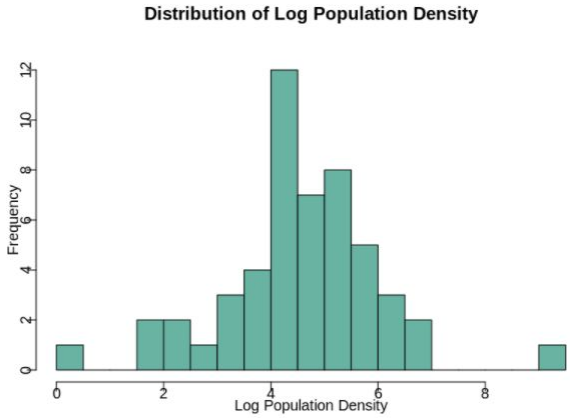
Duration of Stay at home & mask mandate



Comments: Some of the states do not have shelter in place or mask mandates which is why the duration is zero for these states.

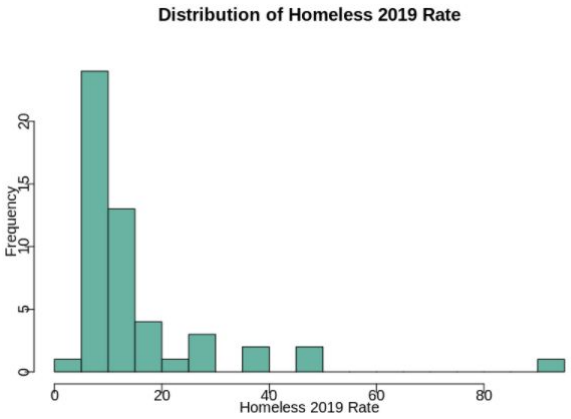
Population Density

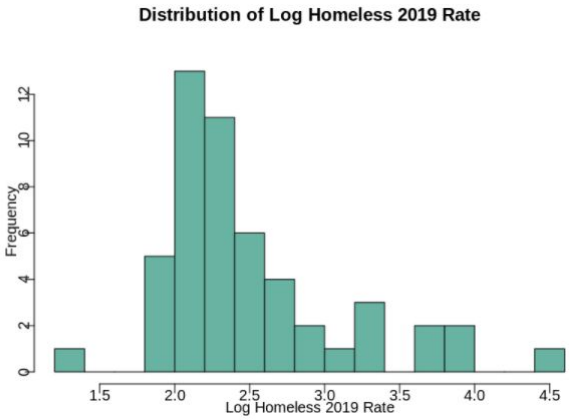




Comments: The population density is skewed to the right. A log of the data makes it close to a normal distribution.

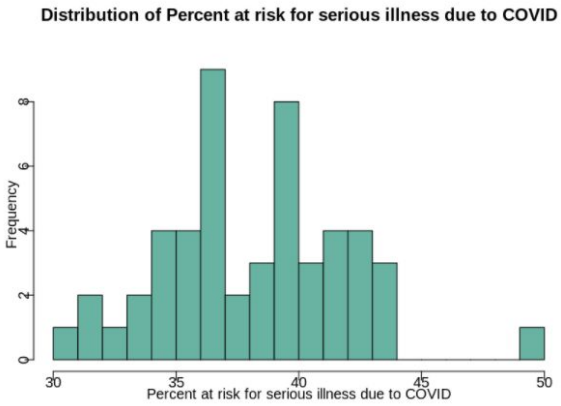
Homeless 2019 Rate

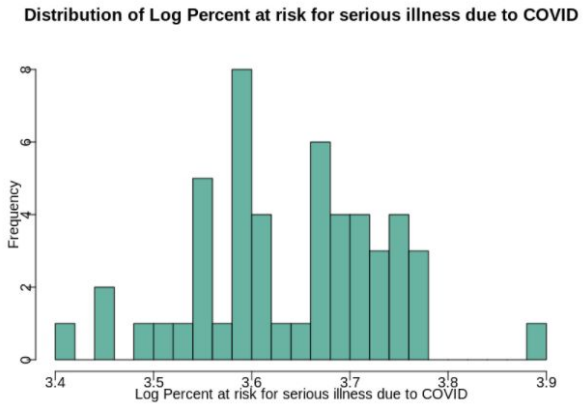




Comments: The homeless rate is skewed to the right. A log of the data makes it closer to a normal distribution.

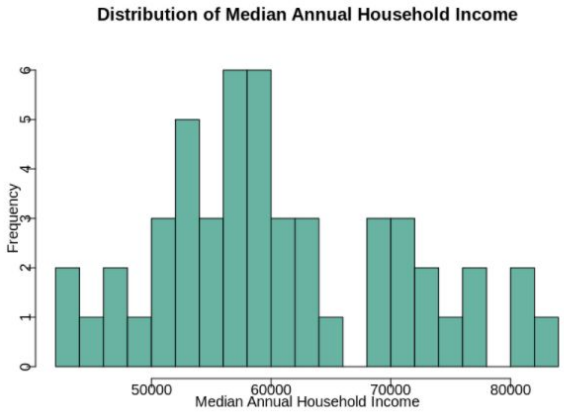
Percent at risk for serious illness due to COVID

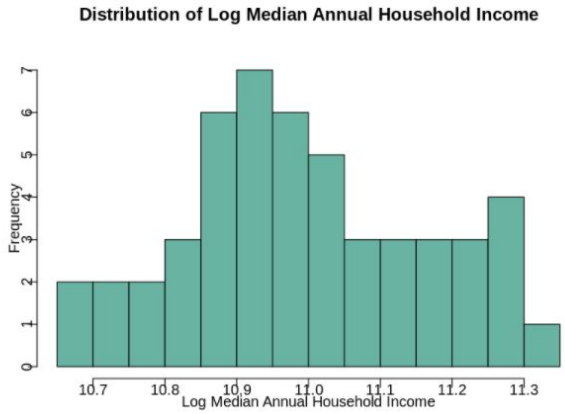




Comments: The distribution is bi-modal and taking log doesn't make any difference in the shape of the distribution.

Median annual household income





Comments: The distribution is close to normal and a log transformation isn't required.

Histograms summary

Based on the above histograms, the distributions of Population Density and Homeless Rate are skewed to the right. It is appropriate to take log transformations of pop_density, and homeless_2019 to make the distribution symmetric. Other variables do not seem to be too skewed and their distributions seem to be fairly symmetric.

3.0 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.

The variables that we want to measure are provided in section 1.1 of this document and listed below:

Cases in Last 7 Days (**infection_last_7_rate**)

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

Our modeling goal is that of explanation.

3. 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?

The covariates that will help achieve our modeling goals are as follows:

Stay at home/ shelter in place (**stay_at_home**)

Mandate face mask use by all individuals in public spaces (**mask_mandate**)

Population density per square miles (**pop_density**)

Percent at risk for serious illness due to COVID (**due_covid_serious_ill_rate**)

Number Homeless (2019) (**homeless_2019_rate**)

Median Annual Household Income (**median_annual_household_income**)

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 will apply **log transformations** for the following covariates:

Population density per square miles

Number Homeless (2019)

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.

Comments: This is discussed in Section 2.0 Exploratory Data Analysis

3.1 Model-1: With Key Variables

We started creating models using key variables first, starting with one independent variable and then adding the remaining key variables.

```
#Model Selection - With Key Variables
model_11=lm(infection_last_7_rate~stay_at_home, data=data, na.action=na.exclude)
coeftest(model_11, vcov = vcovHC)

model_12=lm(infection_last_7_rate~mask_mandate, data=data, na.action=na.exclude)
coeftest(model_12, vcov = vcovHC)

model_13=lm(infection_last_7_rate~legal_enforcement_mask, data=data, na.action=na.exclude)
coeftest(model_13, vcov = vcovHC)

model_14=lm(infection_last_7_rate~stay_at_home+mask_mandate, data=data, na.action=na.exclude)
coeftest(model_14, vcov = vcovHC)

get_robust_se <- function(model) {
  rse <- sqrt(diag(vcovHC(model)))
  return(rse)
}
```

```
t test of coefficients:

             Estimate Std. Error t value Pr(>|t|)
(Intercept)   44.7788     9.1374  4.9006 1.28e-05 ***
stay_at_home -27.1291     9.3222 -2.9102 0.005598 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

t test of coefficients:

             Estimate Std. Error t value  Pr(>|t|)
(Intercept)   28.9326     5.9499  4.8627 1.285e-05 ***
mask_mandate -10.6547     6.2455 -1.7060   0.09447 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

t test of coefficients:

                      Estimate Std. Error t value Pr(>|t|)
(Intercept)            24.4678     3.6012  6.7943 1.52e-08 ***
legal_enforcement_mask  -6.4182     4.4993 -1.4265   0.1602
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

t test of coefficients:

             Estimate Std. Error t value  Pr(>|t|)
(Intercept)   46.8544     9.3934  4.9880 1.003e-05 ***
stay_at_home -25.2309     8.7136 -2.8956  0.005873 **
mask_mandate  -5.5351     4.8945 -1.1309  0.264230
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Stargazer

```
#Review models using Stargazer - With Key Variables
stargazer(model_11,model_12,model_13,model_14,
          type="text",
          se = list(get_robust_se(model_11),get_robust_se(model_12),get_robust_se(model_13),
          get_robust_se(model_14)),
          column.labels = c("Stay at home","Mask Mandate","Legal Enforcement","Final"),
          star.cutoffs = c(0.05, 0.01, 0.001),
          title = "Table 1: Which Policys Best Explain the Change in Infection Rate"
          )
```

```
Table 1: Which Policys Best Explain the Change in Infection Rate
=================================================================================================
                                            Dependent variable:
                    ---------------------------------------------------------------------------
                                            infection_last_7_rate
                        Stay at home         Mask Mandate     Legal Enforcement          Final
                            (1)                  (2)               (3)                   (4)
-------------------------------------------------------------------------------------------------
stay_at_home              -27.129**                                                    -25.231**
                           (9.322)                                                     (8.714)

mask_mandate                                   -10.655                                 -5.535
                                               (6.245)                                 (4.894)

legal_enforcement_mask                                          -6.418
                                                               (4.499)

Constant                  44.779***            28.933***        24.468***            46.854***
                           (9.137)              (5.950)          (3.601)              (9.393)

-------------------------------------------------------------------------------------------------
Observations                 47                   50               50                   47
R2                          0.351                0.091            0.035                0.373
Adjusted R2                 0.337                0.072            0.015                0.344
Residual Std. Error    14.156 (df = 45)     16.300 (df = 48)    16.791 (df = 48)     14.076 (df = 44)
F Statistic           24.382*** (df = 1; 45) 4.794* (df = 1; 48) 1.753 (df = 1; 48) 13.085*** (df = 2; 44)
=================================================================================================
Note:                                                                 *p<0.05; **p<0.01; ***p<0.001
```

In our causal theory, we want to look at the effect of different states' policies on the recent weekly infection rate. Within the limit of the data, we have 4 policies: **stay_at_home, close_business, mask_mandate, legal_enforcement_of_mask_mandate**. From EDA, we already know that we cannot use close business, because 50 out of 51 States had a close business policy, therefore there is not enough degrees of freedom to estimate the effect of close business. Mask mandate and its legal enforcement are really part of the same policy; therefore, we have to pick one or the other in order to correctly measure the impact of the policy. Looking at their p-value for their coefficients, we see that mask mandate has a more significant coefficient. Therefore, in our final version of the **Model-1**, we will include both stay at home and mask mandate and measure their effect on the recent weekly infection rate.
Notice here that the beta for mask mandate is insignificant even though mask has scientifically proven to be effective. We will discuss why the beta is insignificant in a later section.

3.2 Model-2: Key Variables and Covariates

In Model-2 we added the covariates to the model along with the key variables.

```
# Model Selection - With Key Variables and Covariates
# key variables (policies)
model_21=lm(infection_last_7_rate~stay_at_home+mask_mandate,
        data=data, na.action=na.exclude
        )
coeftest(model_21, vcov = vcovHC)

#add serious ill rate
model_22=lm(infection_last_7_rate~stay_at_home+mask_mandate+due_covid_serious_ill_rate,
        data=data, na.action=na.exclude
        )
coeftest(model_22, vcov = vcovHC)

#add population density
model_23=lm(infection_last_7_rate~stay_at_home+mask_mandate+due_covid_serious_ill_rate+
        log(pop_density),
        data=data, na.action=na.exclude
        )
coeftest(model_23, vcov = vcovHC)

#add homeless_2019_rate & median_annual_household_income
model_24=lm(infection_last_7_rate~stay_at_home+mask_mandate+due_covid_serious_ill_rate+
        log(pop_density)+log(homeless_2019_rate)+median_annual_household_income,
        data=data, na.action=na.exclude
        )
coeftest(model_24, vcov = vcovHC)
```

```
t test of coefficients:

             Estimate Std. Error t value  Pr(>|t|)
(Intercept)   46.8544     9.3934  4.9880 1.003e-05 ***
stay_at_home -25.2309     8.7136 -2.8956  0.005873 **
mask_mandate  -5.5351     4.8945 -1.1309  0.264230
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

t test of coefficients:

                           Estimate Std. Error t value Pr(>|t|)
(Intercept)                 87.17036   24.99534  3.4875 0.001137 **
stay_at_home               -23.35119    7.84743 -2.9756 0.004783 **
mask_mandate                -6.36479    4.74626 -1.3410 0.186958
due_covid_serious_ill_rate  -1.07793    0.54437 -1.9801 0.054110 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
t test of coefficients:

                           Estimate Std. Error t value  Pr(>|t|)
(Intercept)                92.86805   22.09967  4.2022 0.0001349 ***
stay_at_home              -18.18419    6.87638 -2.6444 0.0114582 *
mask_mandate               -4.04084    4.03056 -1.0026 0.3218186
due_covid_serious_ill_rate -0.84981    0.44269 -1.9196 0.0617122 .
log(pop_density)           -4.59232    1.26887 -3.6192 0.0007880 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

t test of coefficients:

                                  Estimate  Std. Error t value  Pr(>|t|)
(Intercept)                     1.5663e+02  3.1963e+01  4.9003 1.715e-05 ***
stay_at_home                   -1.3444e+01  7.1665e+00 -1.8759 0.0681643 .
mask_mandate                   -3.6068e+00  3.8807e+00 -0.9294 0.3583831
due_covid_serious_ill_rate     -1.3659e+00  5.2111e-01 -2.6211 0.0124312 *
log(pop_density)               -5.4896e+00  1.2979e+00 -4.2298 0.0001369 ***
log(homeless_2019_rate)        -1.1100e+01  3.3399e+00 -3.3234 0.0019418 **
median_annual_household_income -2.7009e-04  1.2824e-04 -2.1061 0.0416826 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```r
#Review models using Stargazer - Second Model
stargazer(model_21,model_22,model_23,model_24,
          type="text",
          se = list(get_robust_se(model_21),get_robust_se(model_22),get_robust_se(model_23),
                    get_robust_se(model_24)),
          column.labels = c("Model-21","Model-22","Model-23","Model-24"),
          star.cutoffs = c(0.05, 0.01, 0.001),
          title = "Table 2: Adding Race, Wealth, Baseline as Controls"
          )
```

Table 2: Adding Race, Wealth, Baseline as Controls

	Dependent variable:			
	infection_last_7_rate			
	Model-21	Model-22	Model-23	Model-24
	(1)	(2)	(3)	(4)
stay_at_home	-25.231**	-23.351**	-18.184**	-13.444
	(8.714)	(7.847)	(6.876)	(7.167)
mask_mandate	-5.535	-6.365	-4.041	-3.607
	(4.894)	(4.746)	(4.031)	(3.881)
due_covid_serious_ill_rate		-1.078*	-0.850	-1.366**
		(0.544)	(0.443)	(0.521)
log(pop_density)			-4.592***	-5.490***
			(1.269)	(1.298)
log(homeless_2019_rate)				-11.100***
				(3.340)
median_annual_household_income				-0.0003*
				(0.0001)
Constant	46.854***	87.170***	92.868***	156.629***
	(9.393)	(24.995)	(22.100)	(31.963)
Observations	47	47	47	46
R2	0.373	0.423	0.525	0.648
Adjusted R2	0.344	0.382	0.480	0.594
Residual Std. Error	14.076 (df = 44)	13.662 (df = 43)	12.539 (df = 42)	11.052 (df = 39)
F Statistic	13.085*** (df = 2; 44)	10.497*** (df = 3; 43)	11.608*** (df = 4; 42)	11.963*** (df = 6; 39)
Note:				*p<0.05; **p<0.01; ***p<0.001

Comments:

In Model-2, we want to introduce a number of controls whose difference among the states affects the outcome variable. Because we have only 48 samples, we decided to limit our parameters to 6 so that we will have enough statistical power to detect significant parameters. We discussed some of the reasons for picking control in the EDA, but we will focus on the important ones here.

Chosen Variables

High Risk Population: the definition of a high risk population from CDC is based on age and preexisting conditions. As high risk population will naturally self quarantine, they will affect our outcome variable

Pop density: People in high population density are more likely to get infected, but they are also more risk averse. We log transform it in our model because it makes the relationship with outcome variables more linear.

Homeless rate: Homeless people are more vulnerable to infection, but as result, people around homeless are also more risk averse. We log transform it in our model because it makes the relationship with outcome variables more linear.

Median annual household income: people with low income works in jobs that are more vulnerable to infection

Excluded Variables

Race because there is no causal link between race and infection rate that is not explained by other factors.
Location mobility because that is the outcome on the right hand side.
Age even though it is relevant because it is already included in high risk population and high risk population produce more statistical significant beta.

Overall Model:

By adding these controls, our beta for our key explanatory variables dropped significantly. **stay at home** dropped from 25 bps to 13 bps is now statistically insignificant. The beta for all the control variables are significant, suggesting they all have a meaningful impact on the infection rate. Finally, our adjusted R^2 also increased from 0.35 to 0.59. Therefore, these controls help us better explain the infection rate and their presence helped us estimate our key explanatory beta better.

Wald test to check for joint significance of policy features

We used Wald test to check if addition of population density to Model-22 resulted in any statistically significant improvement in Model-23. Reviewing the p-value (0.0007879577) we reject the null hypothesis that addition of this feature does not improve Model-22. We decided to add population density in Model-24.

```
# Test to check if addition of policy related features are jointly significant
waldtest(model_22, model_23, vcov = vcovHC)
```

A anova: 2 × 4

	Res.Df	Df	F	Pr(>F)
	<dbl>	<dbl>	<dbl>	<dbl>
1	43	NA	NA	NA
2	42	1	13.09878	0.0007879577

3.3 Model-3: Key Variables and many Covariates

In Model-3 we added more covariates to the model created in the previous section.

```
# Model 3
#add race, age, and test rate
model_31=lm(infection_last_7_rate~stay_at_home+mask_mandate+due_covid_serious_ill_rate+
        log(pop_density)+log(homeless_2019_rate)+median_annual_household_income+
        log(poverty)+white_pop_rate+black_pop_rate+hispanic_pop_rate+
        pop_under18+pop_19_25+pop_26_34+pop_35_54+pop_54_64,
        data=data, na.action=na.exclude
        )
coeftest(model_31, vcov = vcovHC)
```

```
t test of coefficients:

                                 Estimate  Std. Error t value Pr(>|t|)
(Intercept)                     2.0032e+02  8.1204e+01  2.4669  0.02001 *
stay_at_home                   -1.1620e+01  7.0320e+00 -1.6524  0.10962
mask_mandate                   -3.0939e+00  5.3948e+00 -0.5735  0.57089
due_covid_serious_ill_rate     -1.5649e+00  6.5733e-01 -2.3807  0.02432 *
log(pop_density)               -6.8720e+00  3.2559e+00 -2.1107  0.04386 *
log(homeless_2019_rate)        -2.3472e+01  9.9259e+00 -2.3648  0.02521 *
median_annual_household_income -2.6238e-04  2.3508e-04 -1.1161  0.27385
log(poverty)                    1.9400e+00  7.0929e+00  0.2735  0.78647
white_pop_rate                 -2.0844e-01  3.3109e-01 -0.6296  0.53408
black_pop_rate                 -5.2421e-01  5.8894e-01 -0.8901  0.38100
hispanic_pop_rate              -3.0992e-01  4.9518e-01 -0.6259  0.53647
pop_under18                    -9.1984e-05  5.3627e-05 -1.7153  0.09735 .
pop_19_25                       8.1459e-05  5.1310e-05  1.5876  0.12361
pop_26_34                       3.7889e-06  4.3954e-05  0.0862  0.93192
pop_35_54                       7.6577e-05  5.7404e-05  1.3340  0.19295
pop_54_64                      -4.5946e-05  4.6077e-05 -0.9971  0.32723
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
#Review models using Stargazer - With Key Variables and many Covariates
stargazer(model_24,model_31,
          type="text",
          se = list(get_robust_se(model_24),get_robust_se(model_31)
                    ),
          column.labels = c("Model-24","Overfitted"),
          star.cutoffs = c(0.05, 0.01, 0.001),
          title = "Table 3: Model-24 vs Overfitted Model"
          )
```

	(1)	(2)
stay_at_home	-13.444	-11.620
	(7.167)	(7.032)
mask_mandate	-3.607	-3.094
	(3.881)	(5.395)
due_covid_serious_ill_rate	-1.366**	-1.565*
	(0.521)	(0.657)
log(pop_density)	-5.490***	-6.872*
	(1.298)	(3.256)
log(homeless_2019_rate)	-11.100***	-23.472*
	(3.340)	(9.926)
median_annual_household_income	-0.0003*	-0.0003
	(0.0001)	(0.0002)
log(poverty)		1.940
		(7.093)
white_pop_rate		-0.208
		(0.331)
black_pop_rate		-0.524
		(0.589)
hispanic_pop_rate		-0.310
		(0.495)
pop_under18		-0.0001
		(0.0001)
pop_19_25		0.0001
		(0.0001)
pop_26_34		0.00000
		(0.00004)
pop_35_54		0.0001
		(0.0001)
pop_54_64		-0.00005
		(0.00005)
Constant	156.629***	200.323*
	(31.963)	(81.204)
Observations	46	44
R2	0.648	0.726
Adjusted R2	0.594	0.579
Residual Std. Error	11.052 (df = 39)	11.234 (df = 28)
F Statistic	11.963*** (df = 6; 39)	4.947*** (df = 15; 28)
Note:		*p<0.05; **p<0.01; ***p<0.001

In model 3, we overfit the model2 with additional control variables like race, age, and test rate. We see that when we overfit the model, both the standard error and p-value increases, making our coefficients insignificant. This is due to less degree of freedom, and less unique variance for each variable. Therefore, overfitting a model actually decreases our statistical power to detect the significant coefficients.

However, overfit does allow us to see whether our coefficients are robust, we see that even after adding these controls our coefficients for stay at home and mask mandate are still pretty similar to before. This gives us more confidence that our estimate for the policy effect size is within a reasonable range, even though it is not statistically significant.

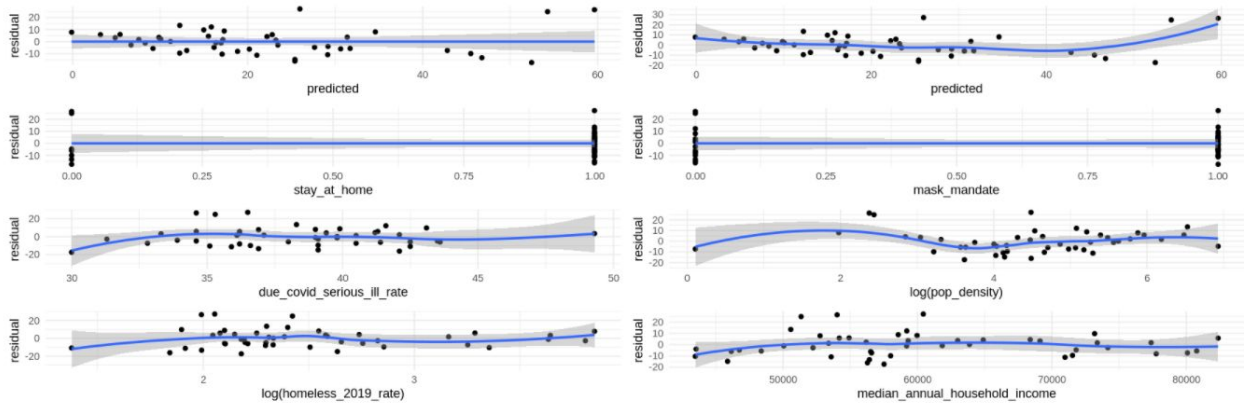
4.0 Limitations of Model

In this section we discuss the limitations of Model_24 which is our final model.

4.1 Assumption-1: Linear population model

Looking at the model's residual against the predictive error and each variable in the model, we see that overall, the relationship is pretty linear except for at extreme ends of the parameters. Furthermore, our key explanatory variables stay at home and mask mandate are very linear, therefore their coefficients are not statistically biased.

```
data <- data %>%
  mutate(residual = resid(model_24),
         predicted = predict(model_24)
        )

plot_1 <- data %>%
  ggplot(aes(x = predicted, y = residual)) +
  geom_point() + geom_smooth(method = 'lm')

plot_2 <- data %>%
  ggplot(aes(x = predicted, y = residual)) +
  geom_point() + stat_smooth()

plot_3 <- data %>%
  ggplot(aes(x = stay_at_home, y = residual )) +
  geom_point() + geom_smooth(method = 'lm')

plot_4 <- data %>%
  ggplot(aes(x = mask_mandate, y = residual)) +
  geom_point() + stat_smooth()

plot_5 <- data %>%
  ggplot(aes(x = due_covid_serious_ill_rate, y = residual)) +
  geom_point() + stat_smooth()

plot_6 <- data %>%
  ggplot(aes(x = log(pop_density), y = residual)) +
  geom_point() + stat_smooth()

plot_7 <- data %>%
  ggplot(aes(x = log(homeless_2019_rate) , y = residual)) +
  geom_point() + stat_smooth()

plot_8 <- data %>%
  ggplot(aes(x = median_annual_household_income , y = residual)) +
  geom_point() + stat_smooth()

(plot_1 | plot_2) /
(plot_3 | plot_4) /
(plot_5 | plot_6) /
(plot_7 | plot_8)
```



4.2 Assumption-2: Random Sampling

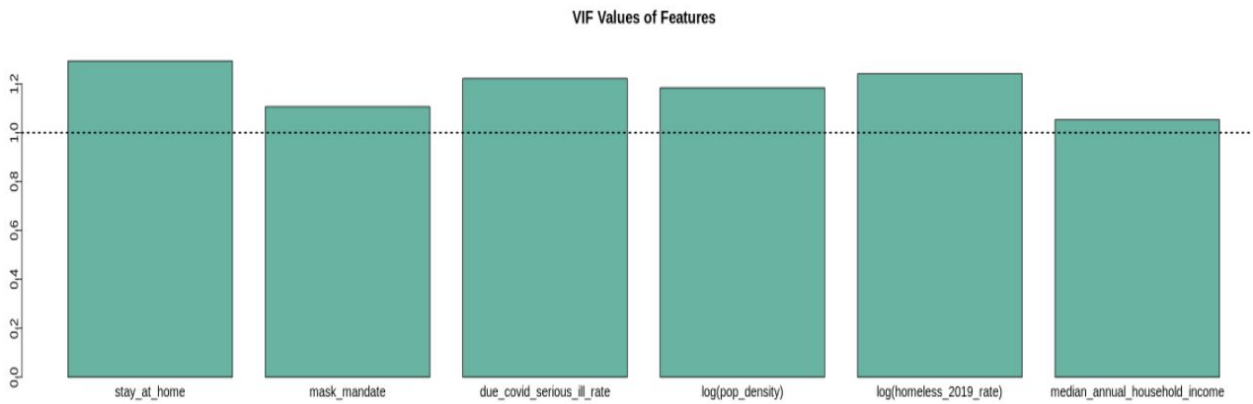
The data is obtained from COVID-19 US state policy database (CUSP) (Raifman J, Nocka K, Jones D, Bor J, Lipson S, Jay J, and Chan P. (2020). " Available at: www.tinyurl.com/statepolicies). The data is for all 50 states in the US and the District of Columbia. Because we have the **entire population** in our data, our data should have the same characteristic as the entire population. However, there are several concerns:

1. Different States may report data differently. This could lead to each sample **not** being identically distributed. If this were the case, our result will not be generalizable to the entire population.
2. Due to the time gap between now and when the metric was collected, some of the data points may no longer be representative. Therefore, we should always take a grain of salt when it comes to a model's interpretations.

4.3 Assumption-3: No perfect multicollinearity

There is no **perfect** nor **near perfect** multicollinearity because all except for one variable is statistically significant. Looking at variance inflation factor (VIF) for each variable, none of the variables has VIF>5, which means that there is very little multicollinearity.

```
# calculate VIF
vif_values <- vif(model_24)
barplot(vif_values, main = "VIF Values of Features", horiz = FALSE, col = "#69b3a2")
abline(h = 1, lwd = 2, lty = 3)
```




```
# chi-square test for multi-colliniarity
tbl = table(data$stay_at_home+data$mask_mandate+data$due_covid_serious_ill_rate+
        data$pop_density+data$homeless_2019_rate+data$median_annual_household_income)
chisq.test(tbl)
```

```
        Chi-squared test for given probabilities

data:  tbl
X-squared = 0, df = 45, p-value = 1
```

The p-value of 1 is greater than the .05 significance level. We accept the null hypothesis which implies that the feature variables are independent.

4.4 Assumption-4: Zero conditional mean

Looking at the residual vs each variable plot from the graphs above, we see that all of the relationships are pretty linear. That means the conditional mean is close to 0.

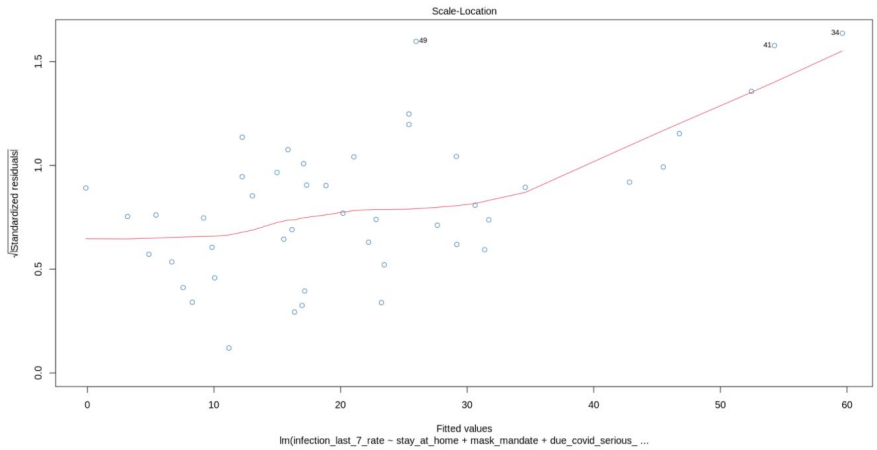
4.5 Assumption-5: Homoscedasticity

Looking at the residual vs predicted plot above, we see that there is a widening of residual on the extreme ends of the predicted value. This suggests that there is heteroskedasticity. Therefore, we used robust standard error rather than classical standard error in our estimation of the standard error to account for the heteroskedasticity.

One of the reasons we might have heteroskedasticity is that there are some other omitted variables or interactions that we are not including that would have helped us better predict the expected infection rate at extreme ends. We will discuss these in the later sections.

The Homoscedasticity assumption can also be checked by examining the Scale-location plot or the spread-location plot as shown below.

```
plot(model_24, 3, col='#4682B4')
```



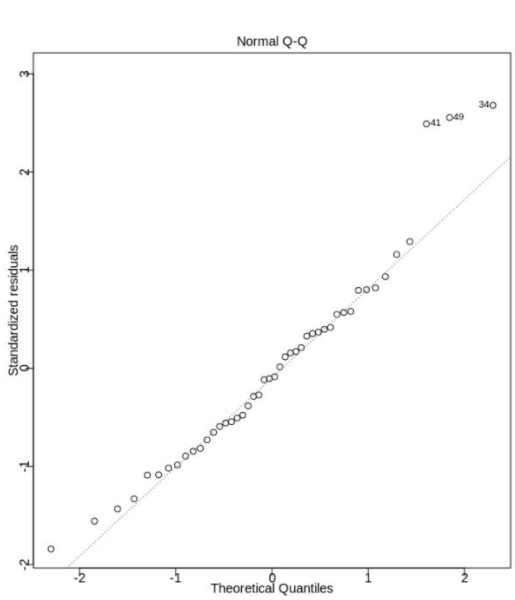
This plot above also shows that the residuals widening on the ends of the predicted value suggesting heteroskedasticity.

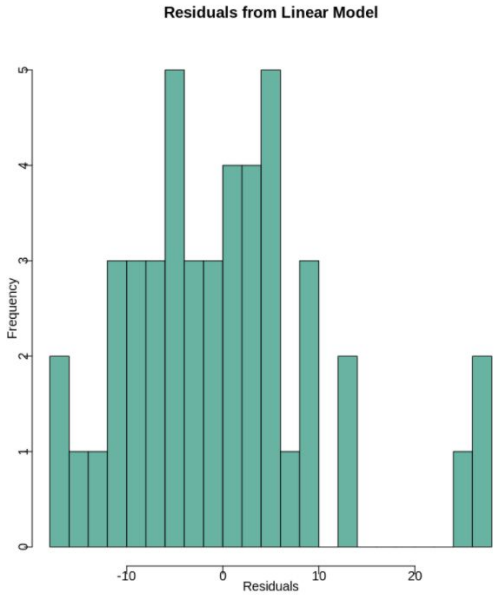
4.6 Assumption-6: Normality of errors

We have used qqplot to review the normality of errors, below. The histograms of residuals are used for visual verification. We see that the residual plot is **only approximately** normally distributed, there is some skewness to the left. Therefore, the P-values we have for the betas in the previous section is a little underreported, which undermines the model conclusion that the coefficient for **stay at home** is statistically significant.

Furthermore, this could suggest there might be some other omitted variables or interactions that we are not including that would have help us better predict the expected infection rate *at extreme ends*. We will discuss these in the later sections.

```
plot(model_24, which = 2)
hist(model_24$residuals, main="Residuals from Linear Model",
    xlab="Residuals", breaks=20, col="#69b3a2")
```





The plot and histogram of the residuals above show that the residuals have a normal distribution. Based on these two graphs we can assume normality of errors.

4.7 Summary of CLM Assumptions

Assumption	Met/Not-Met	Action Taken
Linear Population Model	Mostly Met	Variable transform
IID Sample	Not Met	Acknowledge in report
Zero conditional mean	Mostly met	Variable transform
No perfect multicollinearity	Met	N/A
Homoscedasticity	Not Met	Use Robust Method
Normality of errors	Mostly Met	Variable transformation

5.0 Regression Table

The approach taken in this research is to start with a single key variable (**stay_at_home, mask_mandate, legal_enforcement_mask**) This is shown earlier in Section 2.1 Model-1 (Models: *Model_11-Model_14*).

We then added covariates that we believed enhanced our model. This is shown in Section 2.2 Model-2 (Models: *Model_21-Model24*)

Finally, we added additional covariates to our model in Section 2.3 Model-3 (Models: *Model_31*).

5.1 Stargazer

For this discussion, we picked Models: **Model_14** (no covariates), **Model_24** (explanatory variables and covariates that we believed advanced our model) and **Model_31** (that included many additional covariates)

```
#Review models using Stargazer - With Key Variables and many Covariates
stargazer(model_14,model_24,model_31,
          type="text",
          se = list(get_robust_se(model_14),get_robust_se(model_24),get_robust_se(model_31)
                    ),
          column.labels = c("Model-14-Key Variables","Model-24-Covariates",
                            "Model-31-additional Covariates"),
          star.cutoffs = c(0.05, 0.01, 0.001),
          title = "Table 4: Regression Table - Key Variables and Covariates"
          )
```

	(1)	(2)	(3)
stay_at_home	-25.231**	-13.444	-11.620
	(8.714)	(7.167)	(7.032)
mask_mandate	-5.535	-3.607	-3.094
	(4.894)	(3.881)	(5.395)
due_covid_serious_ill_rate		-1.366**	-1.565*
		(0.521)	(0.657)
log(pop_density)		-5.490***	-6.872*
		(1.298)	(3.256)
log(homeless_2019_rate)		-11.100***	-23.472*
		(3.340)	(9.926)
median_annual_household_income		-0.0003*	-0.0003
		(0.0001)	(0.0002)
log(poverty)			1.940
			(7.093)
white_pop_rate			-0.208
			(0.331)
black_pop_rate			-0.524
			(0.589)
hispanic_pop_rate			-0.310
			(0.495)
pop_under18			-0.0001
			(0.0001)
pop_19_25			0.0001
			(0.0001)
pop_26_34			0.00000
			(0.00004)
pop_35_54			0.0001
			(0.0001)
pop_54_64			-0.00005
			(0.00005)
Constant	46.854***	156.629***	200.323*
	(9.393)	(31.963)	(81.204)
Observations	47	46	44
R2	0.373	0.648	0.726
Adjusted R2	0.344	0.594	0.579
Residual Std. Error	14.076 (df = 44)	11.052 (df = 39)	11.234 (df = 28)
F Statistic	13.085*** (df = 2; 44)	11.963*** (df = 6; 39)	4.947*** (df = 15; 28)
Note:			*p<0.05; **p<0.01; ***p<0.001

Comments:

Putting the models side by side, we see model 2 is better in both adjusted R^2 and F statistics, it suggests that we added the right variables to the model. When we overfit the model with race

and demographic information, although the R^2 went up, its adjusted R^2 and F statistic both went down, suggesting that a lot of the newly added variables don't have much unique information. Furthermore, adding these variables decreased the degree of freedom so much that some of the variables are statistically insignificant any more. Even with overfitting, we see population density and homeless rate and high risk population percentage are still statistically significant. This tells us that these 3 variables are very significant in explaining the infection rates.

5.2 Statistical Significance

1. **due_covid_serious_ill_rate**: This feature represents health risks of people susceptible to serious illness. This is statistically significant. The coefficient is negative meaning more health-risk people leads to less infection rate. One reason why this can happen, is that this category of people is more cautious in their lifestyle to prevent serious illness.

2. **log(pop_density)**: Population density is a statistically significant variable in our models. The coefficient is negative but counter-intuitive as it means a more densely populated area leads to less infection rate. One reason why this can happen, is that people living in highly populated areas are more cautious in their lifestyle to prevent serious illness. Unfortunately cautiousness is an omitted variable we didn't include in the model, this may have caused the beta to become **negative.**

3. **log(homeless_2019_rate)**: This feature captures the socio-economic factors that might influence the infection rate. The Homeless are not able to shelter unless housed by the state (eg. cities using hotels for long term housing). Thus, this should positively correlate with the infection rate. However, the coefficient is negative. This is counterintuitive as it means a more homeless individual in the area leads to less infection rate. One reason arguably is that cities with large homeless populations already have structures for large social services that may be used to combat infection. Unfortunately cautiousness is an omitted variable we did not include in the model, this may have caused the beta to become **negative.**

4. **median_annual_household_income**: This feature also captures if socio-economic status of the population has an impact on the infection rate. For people in the lower income group, it is difficult to adhere to stay at home policies or work from home. As expected it is negatively correlated with the infection rate. This feature seems to be statistically significant.

5. **mask_mandate**: This feature checks if the state has mandated a mask policy. As expected, this feature is negatively correlated with the infection rate. This feature is statistically insignificant.

6. **stay_at_home**: This feature checks if the state has mandated a stay at home policy. As expected, this feature is negatively correlated with the infection rate. This feature is statistically insignificant.

5.3 Practical Significance

1. **due_covid_serious_ill_rate**: From Model-24, we have 1.4 basis point decrease in infection rate for a percent increase in at risk population. This is marginally significant, but it does suggest health and population demographic plays an important role in the infection rate.

2. **log(pop_density)**: From Model-24, we have 5.5 basis point decrease in infection rate if there is a 1% increase in population density. This seems marginally practically significant, considering the skewed population density across different states. This is counter intuitive and requires further research. One possibility is that States with high population density rates take the pandemic more seriously and enact policy earlier. As a result, we cannot get much information from this beta, other than there is some omitted variable.

3. **log(homeless_2019_rate)**: From Model-24, 1% increase in homeless population rate decreases infection rate by 11 basis points. This is counter intuitive and requires further research. One possibility is that States with high homeless rates have more experience deploying and utilizing health and social services. As a result, we cannot get much information from this beta, other than there is some omitted variable.

4. **median_annual_household_income**: From Model-24, 10,000 increase in median annual house income is associated with 3 bps less weekly infection rate. Although it is not a lot given the cost, it does suggest income plays a role in infection.

5. **mask_mandate**: This feature is practically significant. A mask mandate reduces weekly infection rate by 4 basis points. If the beta is true (we don't know because it is not statistically significant), the mask mandate has more effect as increasing annual household income by 10,000 that is pretty significant.

6. **stay_at_home**: This feature is practically significant as stay at home policy reduces the weekly infection rate by 13.444 basis points.

6.0 Omitted Variables

In this section we discuss the effect of omitted variables for Model_24.

In this lab, we tried to model the Covid-19 infection rate last 7 days for each state. In the best model (Model_24), we considered the following factors:

1. **Demography factors**: Population density in each state, percent population susceptible to serious illness as COVID-19 is a highly infectious disease

2. **Socio-Economic factors**: Homeless rate, median annual household income as these factors may determine their behavior during the pandemic.

These factors provide a board picture of the COVID-19 crisis. There are other features in the data set that may provide a better understanding of the overall infection rate due to COVID-19. We feel the following omitted variables if included can lead to improvement of our model.

6.1 Infection rate before policies take place

The first omitted variable that we looked into is: *infection rate before policies take place*

The infection rate before policy takes place has effect on both the state COVID-19 policies and recent weekly infection. The higher the infection rate before the policies take place, the more likely for the State government to enact these policies (positive correlation). Furthermore, higher the infection rate before policies take place, higher the recent infection rate (positive correlation). Because this omitted variable has a positive correlation with both the explanatory variable and the outcome variable, the direction of the bias is also positive. Therefore, because we did not include it in our model, the betas for our state policies (**stay_at_home and mask_mandate**) are less negative than it should be and biased towards 0. It may partially explain why we fail to detect a significant beta for mask mandate and stay at home.

6.2 % of the population whose belief that personal liberty is more important than health (cautiousness or attitude toward covid)

The first omitted variable that we looked into is: *% of the population whose belief that personal liberty is more important than health*

If a State has enough people believe that personal liberty is more important than health then the State is less likely to enact COVID-19 policies (negative correlation). Furthermore, people in these States are more likely to ignore policies if put in place (positive correlation). Because this omitted variable infection rate before policies take place has a negative correlation with the explanatory variable but positive correlation with the outcome variable, the direction of the bias is negative. Therefore, because we did not include it in our model, the betas for our state policies (**stay_at_home and mask_mandate**) are more negative than it should be and biased away from 0. This may also explain why high homeless rates and high population density has negative coefficients.

Fortunately, because this bias has the opposite direction compared to the bias from the last omitted variable, in combination, the total bias is reduced, making the betas for our model more credible.

7.0 Conclusion

We failed to detect a statistical effect size for **stay at home and mask mandate**, even though both should reduce human contact and infection rate. There are 5 possible reasons that may explain this counter intuitive conclusion.

- Wearing a mask over amplifies the sense of safety and makes people do things that increase the likelihood of infection.
- Because we didn't include the baseline infection rate into our model, it is likely that States with higher baseline infection rates also are more likely to have a mask mandate. This bias towards zero may have significantly reduced the absolute size of the beta.
- Because mask mandate sometimes overlaps with stay at home order, some of its effect is being included in the beta of stay at home.
- With only 50 States and the District of Columbia, we do not have enough sample size to detect the effectiveness of mask mandate.
- It is difficult to measure the level of adherence to state policies curbing Covid-19 like social distancing, and mask mandate policies. It is difficult to control big crowd behavior as we have seen in various states during protests and beach openings. This is inversely correlated to population density and also inversely correlated to social economic situation (homeless_2019_rate, median_annual_household_income). As a result, the beta coefficients for these features are overestimated. The beta coefficient of percent population susceptible to serious illness (due_covid_serious_ill_rate) is underestimated.

8.0 References

The data is provided in a file within this repository. Majid Maki-Nayeri compiled the data, drawing many variables from the COVID-19 US state policy database (Raifman J, Nocka K, Jones D, Bor J, Lipson S, Jay J, and Chan P.).

Supporting documents like the specific legal language used by states, additional data sources, and much more are available in unstructured format here (https://tinyurl.com/statepolicysources).

www.tinyurl.com/statepolicies