Kumar_Robertson_Maini_Lab3_W203_DRAFT_2

July 19, 2020

1 Understanding the influence of public policy decisions on Covid19 infections across US states

1.1 Introduction

Covid19 has presented a challenge at global scale not experienced in last 100 years. This pandemic has made significant changes in the way we interact with each other and engage in commerce. US lead the world in the number of Covid19 infections and covid related deaths.

US response to this pandemic was different from most of the world in one fundamental way. Whereas most countries in the world had federal government leading state and local government in terms of setting up and enforcing policies, US took the opposite approach. The response to the pandemic was orchestrated by state government and thus there is a huge variation in the public policy decision timelines across states.

US is a very diverse country and states are very different from each other from the perspective of population density, population distribution and other fundamental factors.

Thus a natural question arises:

Can we attribute the variation in state level public policy decisions (specifically the duration of shelter in place and duration of non-essesntial business closure) to the variation in infections per capita after accounting for the fundamental differentiators like population density and susceptible population percentage?

This is an important question towards undersatnding the effect of public policy decisions on the state of infections per state. It is important to state that the public policy features independently might explain a lot of variation in the infection rates in different state. But if we really want to measure the effect of public policy actions, we need to account for the legitimate difference that exist across states. Given COVID19 is a highly contagious disease, polulation density is an obvious affecting factor which might make the difference.

In our subsequent analysis, we chose to do the following:

1. Dependent Variable: We chose $RatePer100000$ as the dependent variable. It is important to note that we are not working with sheer number of infections rather the infections normalized by population size as we expect this to be more correct approach for our exploration
2. Independent Variables: We are planning to work with the folloing dependent variables:
    1. **Population density per square mile**: Variable type (real numbber); we expect population density to be one of the important predictors of the infection rate given Covid19 is an infectious disease

2. **Percent at risk for serious illness due to covid**: Variable type (real number between 0-100); We expect that the next most important aspect is the percent of susceptible population for this highly contagious disease.
3. **Percent living under federal poverty line**: Variable type (real number between 0-100); Although a sub-section of US workforce (specially at the upper echelon of pay-scale) was able to work from home, workforce at the lower pay-scale had to disproportionately take the risk and potentially expose themselves to the risky situation. Hence, we expect this feature to be important when comparing the infection rates across states
4. **Duration of shelter in place order in days (as of 2020-07-06)**: Variable type (Integer); The first policy decision that we examine is the shelter in place order duration across states. We want to check if, after controlling for first 3 variables, we see a statistically significant impact of shelter in place duration on the infection rate
5. **Duration of mandated face-masks in days (as of 2020-07-06)**: Variable type (Integer); The Second policy decision that we examine is the mandated face-mask duration across states. We want to check if, after controlling for first 3 variables, we see a statistically significant impact of this policy on the infection rate

1.1.1 Considerations:

1. Independent variable 2 and 3 might be correlated
2. The public policies that we want to examine might be highly correlated with each other

1.1.2 Potential issues:

1. There might be important variables missing from the dataset which might affect our conclusions. We will discuss thsi more in later sections
2. Poor population size might have significantly changed since 2018 as we don't have the most up to date data on it

It will be important to analyse the independent variable relationships to do this right.

1.2 Exploratory Data Analysis

1.2.1 1. Extracting relevant data

Starting with the large data file, we first will focus on extracting the key features that we have discussed earlier to be part of our consideration set.

[19]:
```
library(zoo)
library(car)
library(lmtest)
library(sandwich)
library(stargazer)
```

[1]:
```
# collecting the dependent and independent variables

y = c('RatePer100000')
X = c('Population.density.per.square.miles',
      'Percent.at.risk.for.serious.illness.due.to.COVID',
      'Percent.living.under.the.federal.poverty.line..2018.',
```

```
      'Stay.at.home..shelter.in.place', 'End.relax.stay.at.home.shelter.in.
 ↪place',
      'Mandate.face.mask.use.by.employees.in.public.facing.businesses')
oth = c('State')

filename = 'Lab3_data.csv'
```

[2]:
```
# Reading the full data
data_full <- read.csv(filename)
```

[3]:
```
# Extracting only relevant columns to the key data frame
data <- data_full[,c(oth, y, X)]
```

[4]:
```
# Changing the column names to be more human readable

colnames(data) <- c('state', 'RatePer100000', 'PopulationDensity',␣
 ↪'PercentAtRisk', 'PercentUnderPoverty',
                    'ShelterInPlaceStart', 'ShelterInPlaceEnd',␣
 ↪'FaceMaskMandated')
```

[5]:
```
# preliminary data summary
summary(data)
```

```
       state    RatePer100000    PopulationDensity  PercentAtRisk
 Alabama   : 1   Min.   :  66.1   Min.   :    1.11   Min.   :30.00
 Alaska    : 1   1st Qu.: 467.5   1st Qu.:   51.69   1st Qu.:35.98
 arizona   : 1   Median : 723.9   Median :   91.11   Median :38.65
 Arizona   : 1   Mean   : 837.3   Mean   :  386.30   Mean   :38.16
 Arkansas  : 1   3rd Qu.:1036.4   3rd Qu.:  204.35   3rd Qu.:40.58
 California: 1   Max.   :4220.4   Max.   :11496.81   Max.   :49.30
 (Other)   :46
 PercentUnderPoverty ShelterInPlaceStart ShelterInPlaceEnd  FaceMaskMandated
 Min.   : 7.60       0        :12        0        :16      0        :10
 1st Qu.:10.97       3/24/2020: 6        5/18/2020: 4      5/1/2020 : 7
 Median :12.85       3/25/2020: 5        5/4/2020 : 4      5/11/2020: 4
 Mean   :12.93       3/28/2020: 5        5/1/2020 : 3      5/4/2020 : 3
 3rd Qu.:14.15       3/30/2020: 4        5/15/2020: 3      5/8/2020 : 3
 Max.   :19.70       3/23/2020: 3        5/29/2020: 3      4/17/2020: 2
                     (Other)  :17        (Other)  :19      (Other)  :23
```

1.2.2 2. Data Cleaning

We see that there are 52 rows of data where each row should represent one US state. Data set included Washignton DC so we expected the data set to contain only 51 rows. After some investigation we find that Arizona has two rows:

[6]:
```
# Extracting data rows for arizona
data[3:4,]
```

	state	RatePer100000	PopulationDensity	PercentAtRisk	PercentUnderPoverty	ShelterInPlaceStar
3	Arizona	1367.7	62.91	39.1	14	3/31/2020
4	arizona	1367.7	62.91	39.1	14	3/31/2020

It seems that both rows have exact same values for all relevant columns. Hence it seems safe to just delete row 4 so that we can have the right dataset.

[7]:
```
data <- data[-c(4),]                # removing row 4
rownames(data) <- 1:nrow(data)      # renaming the rows of dataframe
summary(data)                       # data summary
```

```
       state    RatePer100000    PopulationDensity  PercentAtRisk
 Alabama    : 1   Min.   :  66.1   Min.   :    1.11   Min.   :30.00
 Alaska     : 1   1st Qu.: 458.6   1st Qu.:   48.66   1st Qu.:35.95
 Arizona    : 1   Median : 717.4   Median :   93.24   Median :38.30
 Arkansas   : 1   Mean   : 826.9   Mean   :  392.64   Mean   :38.15
 California: 1   3rd Qu.:1013.2   3rd Qu.:  209.56   3rd Qu.:40.65
 Colorado   : 1   Max.   :4220.4   Max.   :11496.81   Max.   :49.30
 (Other)    :45
 PercentUnderPoverty ShelterInPlaceStart ShelterInPlaceEnd  FaceMaskMandated
 Min.   : 7.60       0        :12        0        :16      0        :10
 1st Qu.:10.95       3/24/2020: 6        5/18/2020: 4      5/1/2020 : 7
 Median :12.80       3/25/2020: 5        5/4/2020 : 4      5/11/2020: 4
 Mean   :12.91       3/28/2020: 5        5/1/2020 : 3      5/4/2020 : 3
 3rd Qu.:14.20       3/30/2020: 4        5/15/2020: 3      4/17/2020: 2
 Max.   :19.70       3/23/2020: 3        5/29/2020: 3      4/18/2020: 2
                     (Other)  :16        (Other)  :18      (Other)  :23
```

[8]:
```
# Converting the date columns to string types
data$ShelterInPlaceStart<- as.character(data$ShelterInPlaceStart)
data$ShelterInPlaceEnd<- as.character(data$ShelterInPlaceEnd)
data$FaceMaskMandated<- as.character(data$FaceMaskMandated)
```

1.2.3 3. Feature Extraction

Reference Date: Our dependent variable is updated as of 2020/07/06, hence we will use this date as the reference date for calculating duration based features

We want to extract the following features from the data: 1. Duration of shelter in place order in days (as of 2020-07-06) 2. Duration of mandated face masks in days (as of 2020-07-06)

For this we need to look out for the following:

- **Replacement of 0 with the reference date (2020/07/06)**:There are a lot of 0s in the date columns. The interpretation of 0 is that the particular policy is either not yet enforeced

(In case of *ShelterInPlaceStart* and *FaceMaskMandated*) or not yet relaxed (in case of *ShelterInPlaceEnd*). We need to do the following:

- Replace all the 0 in the date columns with string '2020/07/06' (i.e. 6th July 2020)
- Convert the date columns to date type
- Calculate the differences (in days) between *ShelterInPlaceStart* and *ShelterInPlaceEnd* column and the date in *FaceMaskMandated* with the refence date
- Check for the negative durations (as this will indicate data quality issues)

[9]:
```
# duration of shelter in place function
sip_func <- function(x){
    ifelse(x[1]=='0', 0, ifelse(x[2]=='0', as.numeric(as.Date('7/6/2020', '%m/
 ↪%d/%Y')-as.Date(x[1], '%m/%d/%Y')), as.numeric(as.Date(x[2], '%m/%d/%Y') -␣
 ↪as.Date(x[1], '%m/%d/%Y'))))
}
```

[10]:
```
# duration of mandated mask
mask_func <- function(x){
    ifelse(x=='0', 0, as.numeric(as.Date('7/6/2020', '%m/%d/%Y')-as.Date(x, '%m/
 ↪%d/%Y')))
}
```

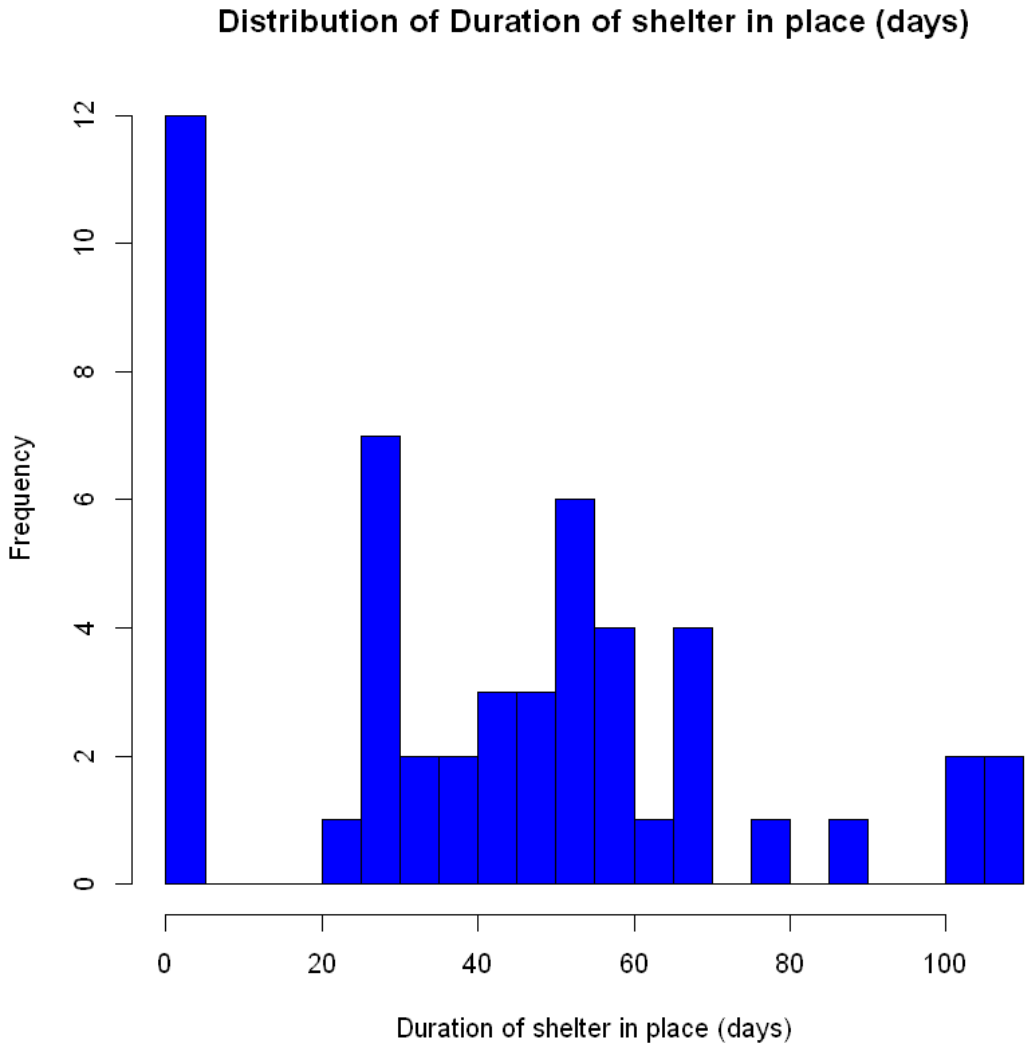
[11]:
```
# Extracting features
data$ShelterInPlaceDuration <- apply(data[, c('ShelterInPlaceStart',␣
 ↪'ShelterInPlaceEnd')], 1, sip_func)
data$MandatedMaskDuration <- apply(data['FaceMaskMandated'], 1, mask_func)
```

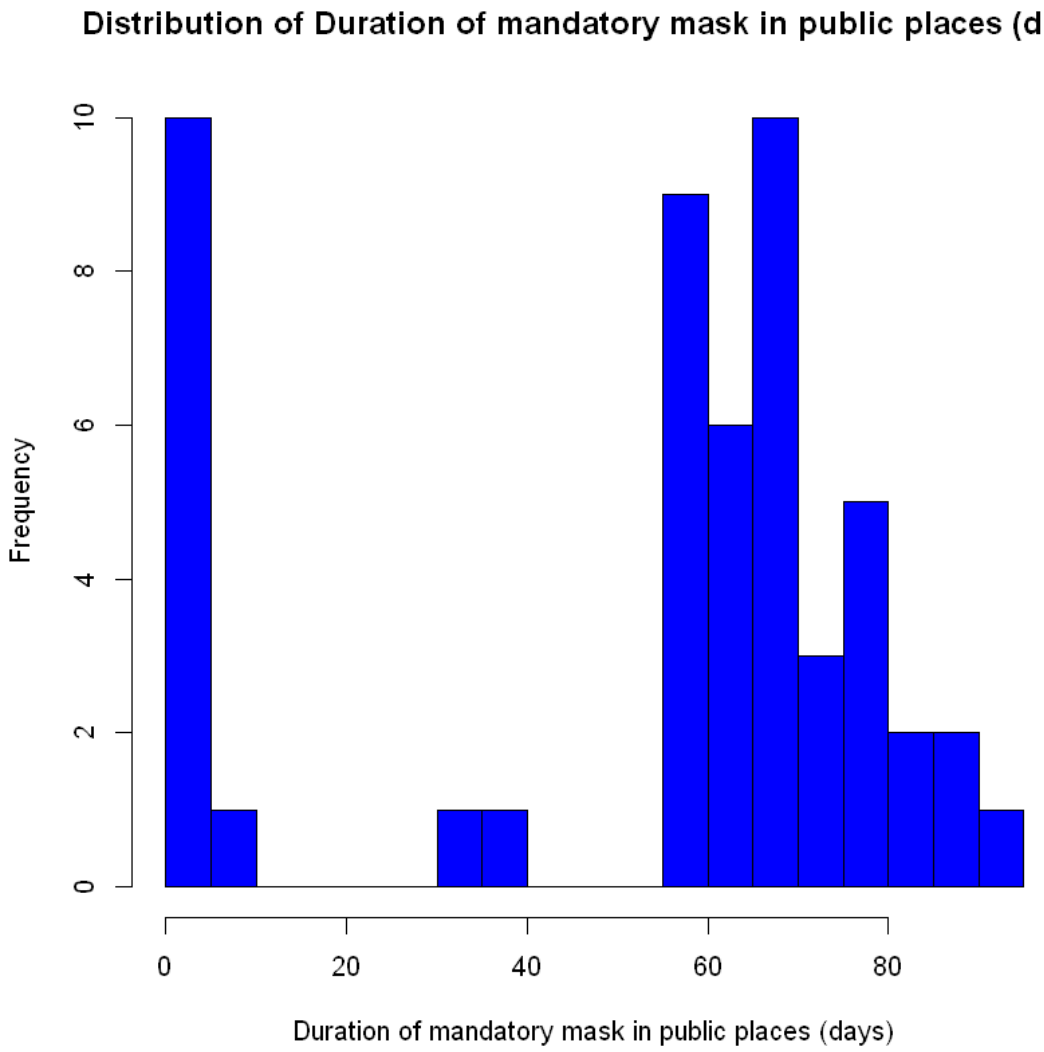
[12]:
```
# Looking at the summary of the new features
summary(data[,c('ShelterInPlaceDuration', 'MandatedMaskDuration')])
```

```
ShelterInPlaceDuration MandatedMaskDuration
Min.   :  0.00         Min.   : 0.00
1st Qu.: 25.00         1st Qu.:47.00
Median : 42.00         Median :63.00
Mean   : 41.45         Mean   :52.75
3rd Qu.: 59.00         3rd Qu.:70.50
Max.   :109.00         Max.   :94.00
```

[13]:
```
# EDA on RatePer100000 , log transform is needed to normalize the dependent␣
 ↪variable.
hist(data$ShelterInPlaceDuration, main = "Distribution of Duration of shelter␣
 ↪in place (days)",
     xlab = "Duration of shelter in place (days)", breaks=20, col='Blue')
hist(data$MandatedMaskDuration, main = "Distribution of Duration of mandatory␣
 ↪mask in public places (days)",
```

```
    xlab = "Duration of mandatory mask in public places (days)", breaks=20,␣
↪col='Blue')
```





The new variables seem not too skewed. Their distribution seem ok. Some of the states don't have policies in place for shelter in place and mandatory mask in public areas, which is resulting in 0 days for these feature. Overall, we are comfortable with these two features.

1.2.4 Cleaning up and extracting only relevant features

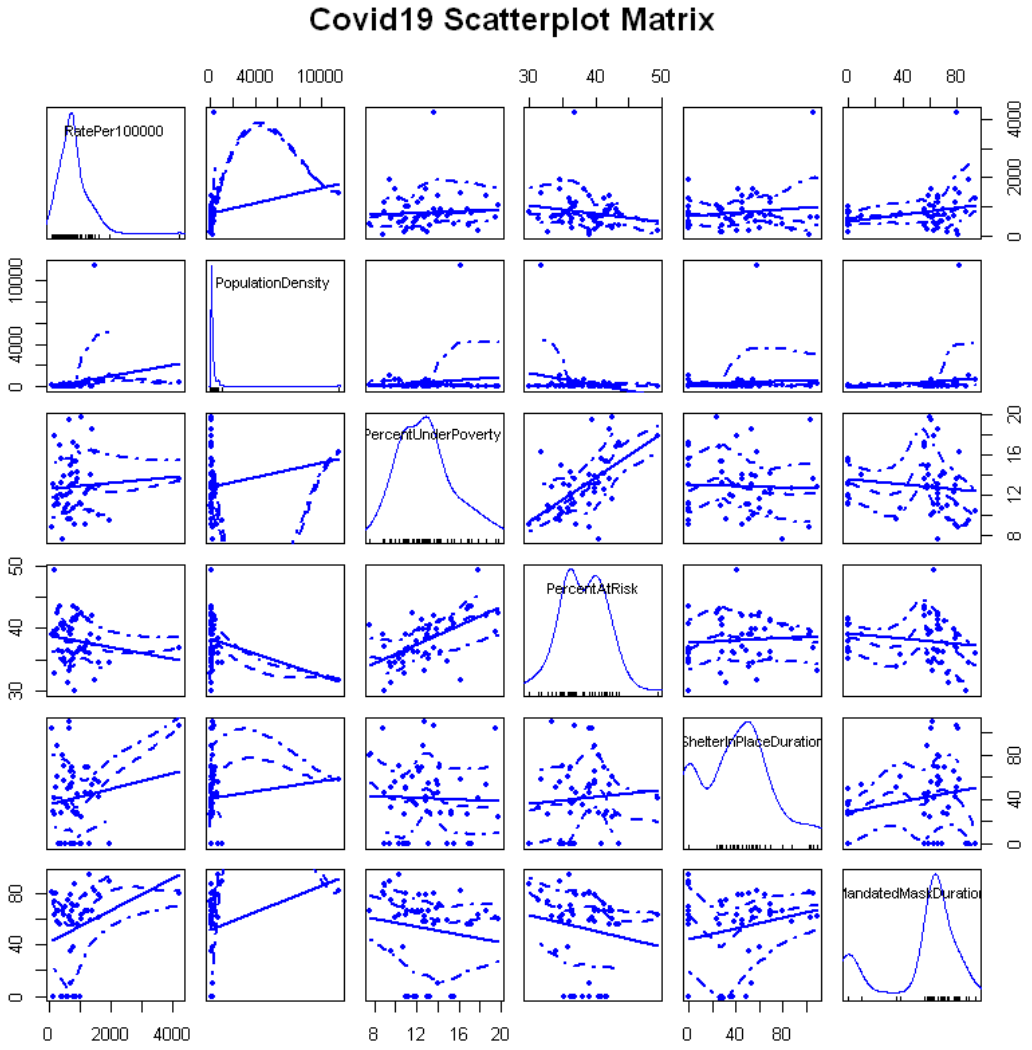
[14]:
```
ref <- c('state')
y <- c('RatePer100000')
x_final <- c('PopulationDensity', 'PercentAtRisk', 'PercentUnderPoverty',␣
 ↪'ShelterInPlaceDuration', 'MandatedMaskDuration')
```

[15]:
```
data <- data[, c(ref, y, x_final)]
```

[16]:
```
summary(data)
```

```
       state    RatePer100000    PopulationDensity  PercentAtRisk
Alabama   : 1   Min.   :  66.1   Min.   :    1.11   Min.   :30.00
Alaska    : 1   1st Qu.: 458.6   1st Qu.:   48.66   1st Qu.:35.95
Arizona   : 1   Median : 717.4   Median :   93.24   Median :38.30
Arkansas  : 1   Mean   : 826.9   Mean   :  392.64   Mean   :38.15
California: 1   3rd Qu.:1013.2   3rd Qu.:  209.56   3rd Qu.:40.65
Colorado  : 1   Max.   :4220.4   Max.   :11496.81   Max.   :49.30
(Other)   :45
PercentUnderPoverty ShelterInPlaceDuration MandatedMaskDuration
Min.   : 7.60       Min.   :  0.00         Min.   : 0.00
1st Qu.:10.95       1st Qu.: 25.00         1st Qu.:47.00
Median :12.80       Median : 42.00         Median :63.00
Mean   :12.91       Mean   : 41.45         Mean   :52.75
3rd Qu.:14.20       3rd Qu.: 59.00         3rd Qu.:70.50
Max.   :19.70       Max.   :109.00         Max.   :94.00
```

[20]:
```
# Plot #2: same as above, but add loess smoother in lower and correlation in␣
 ↪upper
scatterplotMatrix(~RatePer100000+PopulationDensity+PercentUnderPoverty+PercentAtRisk+ShelterIn
      data=data, pch=20, main="Covid19 Scatterplot Matrix")
```



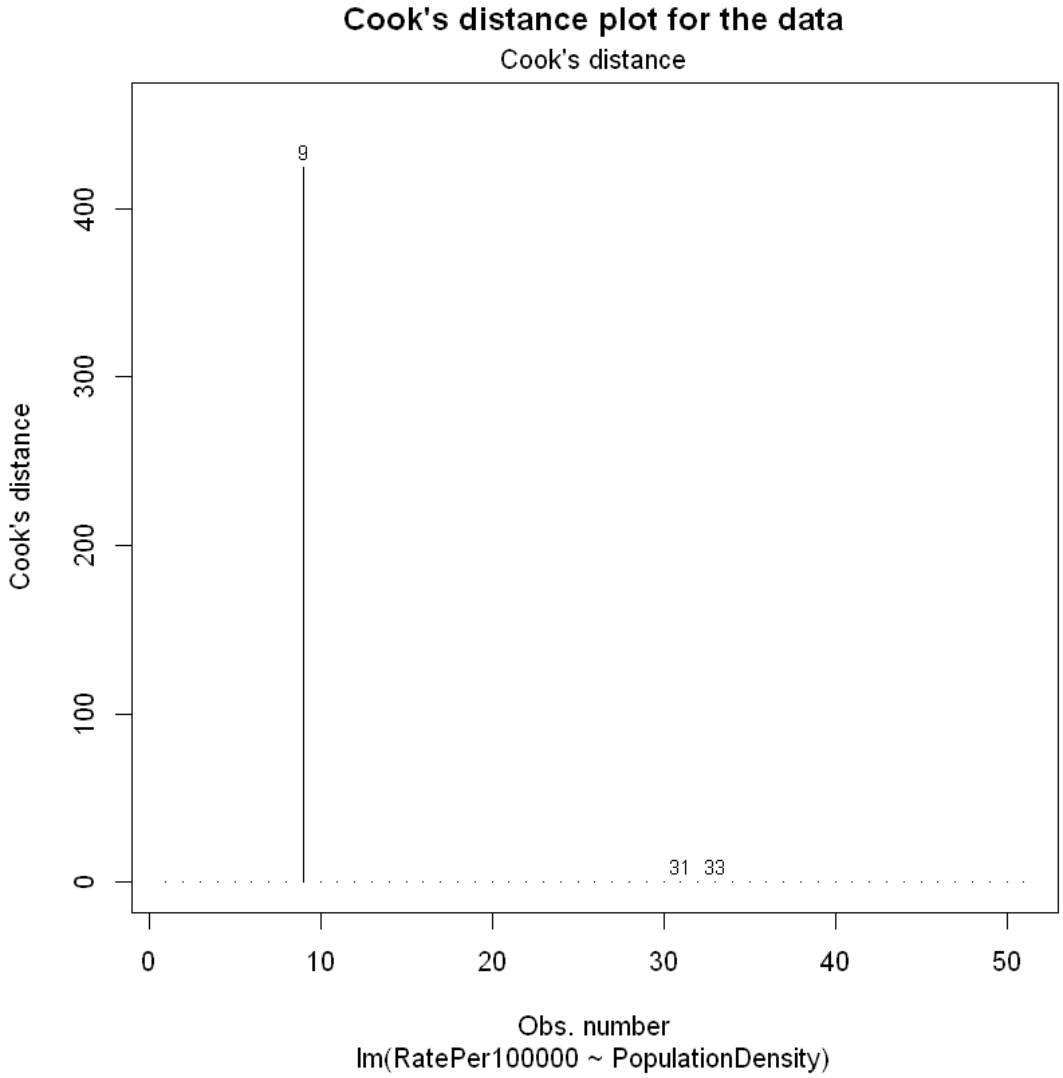
The whole data seems to be skewed by the inclusion of Washigton DC and its huge population density. We can see from the cook distance plot that row 9 (i.e. DC) has a lot of influence on the infection rate in the univariate regression setting.

1. We are comparing states and DC is not a state (and significantly different than a state)
2. From cook's distance analysis, it is clear that DC population density is an outlier

We suggest to remove DC from the dataset.

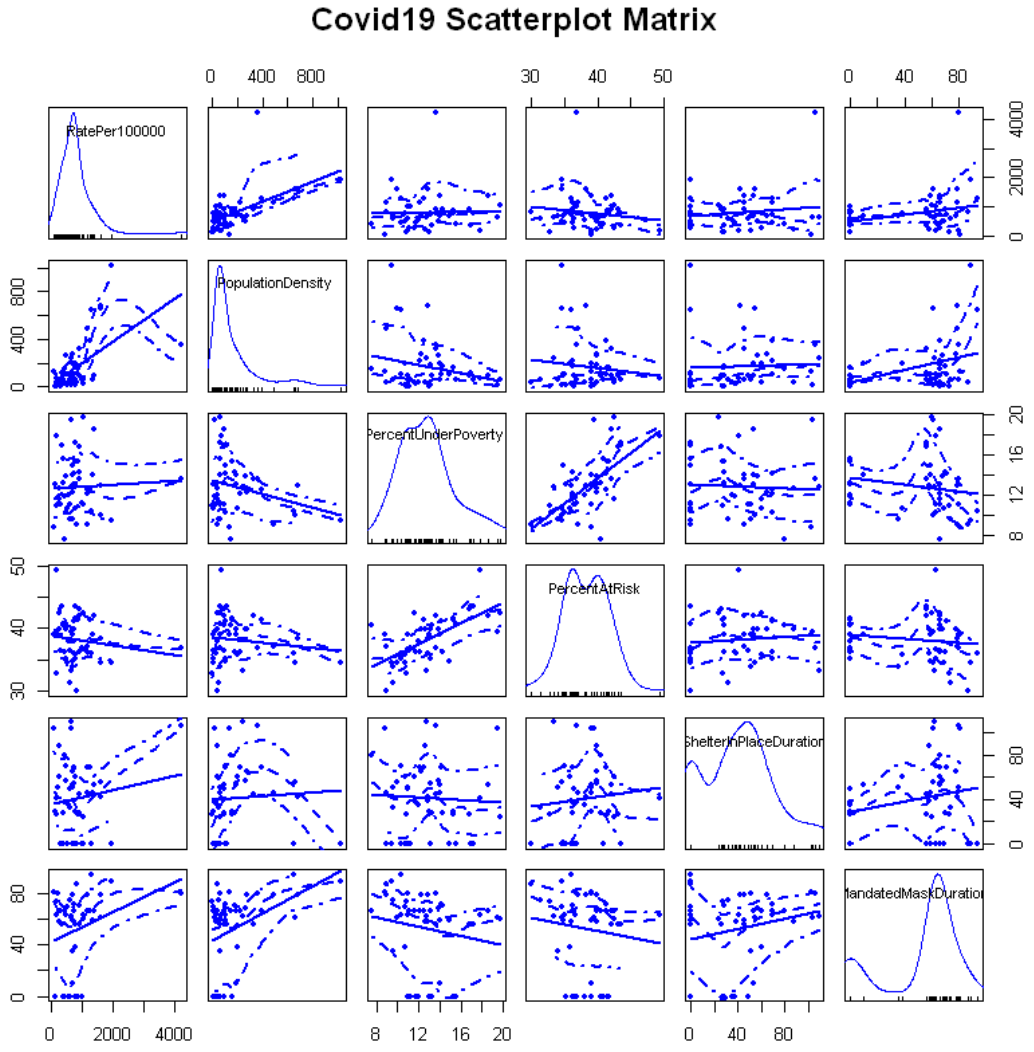
[21]:
```
library(car)

fit <- lm(RatePer100000 ~ PopulationDensity, data=data)
cutoff <- 4/((nrow(data)-length(fit$coefficients)-2))
```

```
plot(fit, which=4, cook.levels=cutoff, main="Cook's distance plot for the␣
 ↪data", )
```



[22]:
```
# Removing DC from the training data
data <- data[!(data$state == 'District of Columbia'),]
```

[23]:
```
scatterplotMatrix(~RatePer100000+PopulationDensity+PercentUnderPoverty+PercentAtRisk+ShelterIn
      data=data, pch=20, main="Covid19 Scatterplot Matrix")
```

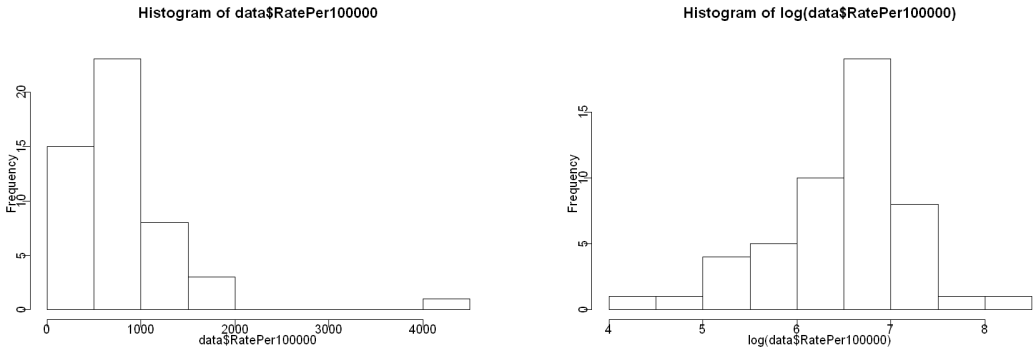


[24]:
```
# Correlation matrix
cor(data[,␣
↪c('RatePer100000','PopulationDensity','PercentAtRisk','PercentUnderPoverty','ShelterInPlace
```

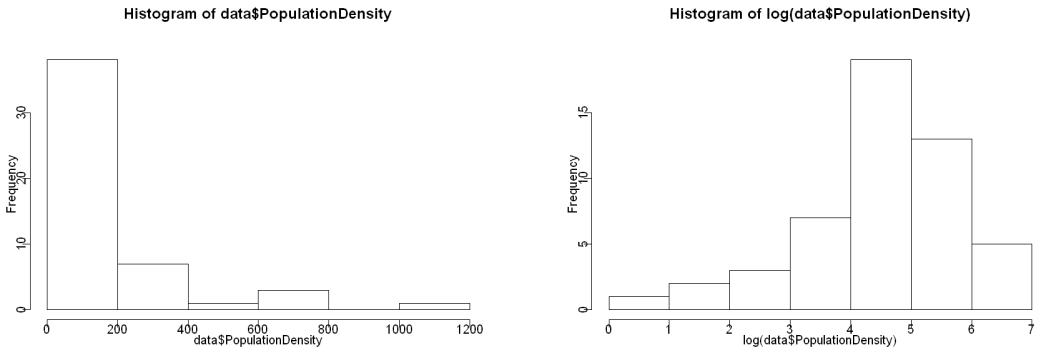
	RatePer100000	PopulationDensity	PercentAtRisk	PercentUnderPoverty	Shelte
RatePer100000	1.00000000	0.55255918	-0.13615223	0.04198269	0.132
PopulationDensity	0.55255918	1.00000000	-0.11974455	-0.24469596	0.052
PercentAtRisk	-0.13615223	-0.11974455	1.00000000	0.64567061	0.096
PercentUnderPoverty	0.04198269	-0.24469596	0.64567061	1.00000000	-0.050
ShelterInPlaceDuration	0.13218571	0.05221688	0.09620189	-0.05019443	1.000
MandatedMaskDuration	0.25127750	0.37411857	-0.12586782	-0.17340814	0.214

1.2.5 Distribution of all independent variables

[65]:
```
# EDA on RatePer100000 , log transform is needed to normalize the dependent␣
↪variable.
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))
hist(data$RatePer100000)
hist(log(data$RatePer100000))
```

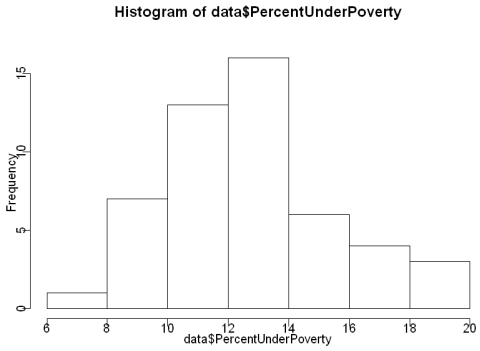


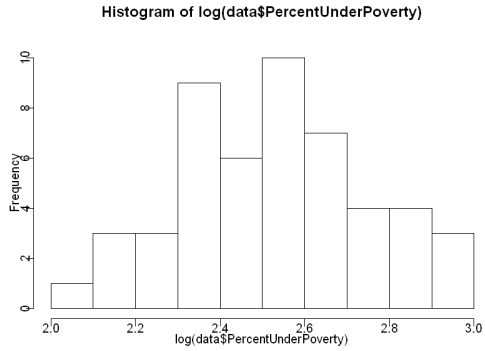
[66]:
```
# EDA on PopulationDensity , log transform is needed to normalize the␣
↪independent variable.
options(repr.plot.height=5, repr.plot.width = 15)
par(mfrow=c(1,2), mar=c(2,2,5,5), mgp=c(.8,.1,0))
hist(data$PopulationDensity)
hist(log(data$PopulationDensity))
```



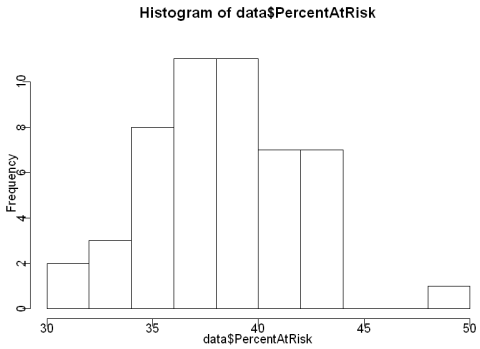
[67]:
```
# EDA on PopulationDensity , log transform is needed to normalize the␣
↪independent variable.
options(repr.plot.height=5, repr.plot.width = 15)
par(mfrow=c(1,2), mar=c(2,2,5,5), mgp=c(.8,.1,0))
```

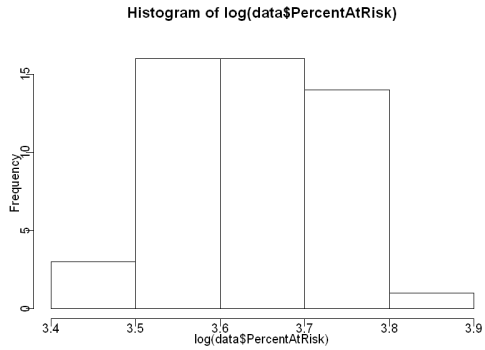
```
hist(data$PercentUnderPoverty)
hist(log(data$PercentUnderPoverty))
```





[68]:
```
# EDA on PopulationDensity , log transform is needed to normalize the␣
 ↪independent variable.
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))
hist(data$PercentAtRisk)
hist(log(data$PercentAtRisk))
```





Based on the above histograms, it is evident that by taking the log transform of **RatePer100000** (i.e. the dependent variable) and **PopulationDensity** (i.e. one of the important independent variable), we can make both of them more symmetric. Remaining two variables are already pretty symmetric. Log transform has a very intuitive impact on the model interpretation as well. Hence we have decided to make this transformation on these two columns.

[25]:
```
data$LogRatePer100000 <- log(data$RatePer100000)
data$LogPopulationDensity <- log(data$PopulationDensity)
```

[26]:
```
write.csv(data, 'final_covid_data.csv', row.names=FALSE)
```

1.3 Model Building Process

1.4 Model types

1.4.1 Baseline Model

Our baseline is very simple model with just 1 independent variable (log(Population_Density) per state). As covid is highly contagious disease, we expect population density to play a key role in detemining the extent of infection in any given state.

1.4.2 Improvement V1 model

Population density is a coarse measure. There are socio economic and health (i.e. pre-existing conditions) related factor that also affect the baseline infection rates. Hence in our candidate model, we add the percent at risk (indicating the percentage of population at risk per state) as our health indicator and percent under poverty as our socio economic indicator. As work from home and comfortable quarentine is incresingly difficult for people with limited means, we expect it to play a big role in determining infection rates.

1.4.3 Improvement V2 model

Finally, we add two more features to our dataset (as described above) both capturing some aspects of public policy decisions that were (or weren't) taken by state authorities. Via this model, we will try to assess if the addition of policy related features has significant effect on the infection rate/100000.

[27]:
```
model1 <- lm(data$LogRatePer100000~data$LogPopulationDensity, data = data)
se.model1 <- sqrt(diag(vcovHC(model1)))

model2 <-␣
↪lm(data$LogRatePer100000~data$LogPopulationDensity+PercentUnderPoverty, data␣
↪= data)
se.model2 <- sqrt(diag(vcovHC(model2)))

model3 <-␣
↪lm(data$LogRatePer100000~data$LogPopulationDensity+PercentUnderPoverty+PercentAtRisk,␣
↪data = data)
se.model3 <- sqrt(diag(vcovHC(model3)))

model4 <-␣
↪lm(data$LogRatePer100000~data$LogPopulationDensity+PercentUnderPoverty+PercentAtRisk+Shelte
↪data = data)
se.model4 <- sqrt(diag(vcovHC(model4)))

model5 <-␣
↪lm(data$LogRatePer100000~data$LogPopulationDensity+PercentUnderPoverty+PercentAtRisk+Shelte
            data = data)
se.model5 <- sqrt(diag(vcovHC(model5)))
```

1.4.4 Wald test for testing joint significance of policy features

We used Wald test to test if the addition of policy features in model 5 resulted in any statistically significant improvement in our V1 model (model 3). Looking at the p value (0.5256), we fail to reject the null hypothesis that addition of these features doesn't improve our v1 model.

[28]:
```
# Testing if the addition if two policy related features are jointly␣
 ↪significant
waldtest(model3, model5, vcov = vcovHC)
```

Res.Df	Df	F	Pr(>F)
46	NA	NA	NA
44	2	0.6526556	0.5256277

1.4.5 Regression table

[41]:
```
stargazer(model1, model2, model3, model4, model5,
          type = "text",
          omit.stat = "f",
          se = list(se.model1, se.model2, se.model3, se.model4, se.model5),
          star.cutoffs = c(0.05, 0.01, 0.001))
```

```
================================================================================
======================
                                                Dependent variable:
-------------------------------------------------------------------------------
                                                   LogRatePer100000
                              (1)             (2)             (3)             (4)
(5)
--------------------------------------------------------------------------------
----------------------
LogPopulationDensity      0.308***        0.313***        0.375***
0.398***        0.397***
                           (0.054)         (0.056)         (0.054)
(0.049)         (0.050)

PercentUnderPoverty                         0.038           0.146**
0.140**         0.140**
                                           (0.038)         (0.052)
(0.052)         (0.052)

PercentAtRisk                                               -0.128***
-0.123**        -0.123**
                                                            (0.038)
(0.039)         (0.040)

ShelterInPlaceDuration
-0.004          -0.004
```

```
(0.004)         (0.004)

MandatedMaskDuration
0.0002
           (0.002)

Constant                   5.086***        4.573***        7.818***
7.760***        7.743***
                            (0.261)         (0.573)         (0.859)
(0.869)         (0.935)

------------------------------------------------------------------------------
----------------------
Observations                  50              50              50               50
50
R2                           0.302           0.323           0.533
0.560           0.560
Adjusted R2                  0.287           0.294           0.502
0.520           0.510
Residual Std. Error    0.633 (df = 48) 0.630 (df = 47) 0.529 (df = 46) 0.519 (df
= 45) 0.525 (df = 44)
==============================================================================
======================
Note:
*p<0.05; **p<0.01; ***p<0.001
```

1.5 Regression Table - Discussion

In the above regression table, we start from a very simple 1-factor model and rpogressively add complexity by adding additional independent variables.

1.5.1 Statistical Significance

1. **LogPopulationDensity**: This is a highly statistically significant independent variable which is consistently significant across all our model choices.
2. **PercentUnderPoverty**: We added this feature to incorporate the socio economic factors that might influence the infection rate. It is harder for people with limited means to remain at their home and do their jobs from home. This factors should be positively correlated to the infection rate. We see that this is a highly significant factor in most of the models it is a part of.
3. **PercentAtRisk**: This factor represents the health risk in each state. We expect this to be important and positively correlated with the infection rate. Although this turns out to be statistically significant in all models it is a part of, the coefficient is negative. This is very counter intuitive as this means that the more vulnerable population a state has, lesser the infection rate is. We discuss more about this in conclusion.
4. **ShelterInPlaceDuration**: This feature calculates the number if days shelter in place rules are enforced. We expect it to be negatively correlated with infection rate. We see that this feature is not statistically significant in any of the models.

5. **MandatedMaskDuration**: This feature calculates the number of days mandatory mask in public place rules are enforced. We expect it to be negatively correlated with infection rate. We see that this feature is not statistically significant in any of the models.

1.5.2 Practical significance

1. **LogPopulationDensity**: From model 3, we have that 1% increase in population density result in 0.375% increase in the infection rate. This seems marginally practically significant as the distribution of population density can be very skewed.

2. **PercentUnderPoverty**: From model 3, we have that 1% increase in Percent under poverty result in 14.6% increase in the infection rate. This is highly practically significant.
3. **PercentAtRisk**: From model 3, we have that 1% increase in Percent at risk result in 12.8% decrease in the infection rate. This is highly practically significant but needs further exploration and explanation.
4. **ShelterInPlaceDuration**: This feature is not practically significant as 1 day increase in shelter in place decreases the infection rate by 0.4%.
5. **MandatedMaskDuration**: This feature is not practically significant as 1 day increase in shelter in place decreases the infection rate by 0.2%.

2 Validation of 6 CLM Assumptions - V1 (Model 3)

2.0.1 Assumption 1: Linear population model:

Linear population model: We don't have to check the linear population model, because we haven't constrained the error term, i.e haven't required it to be normal, so there's nothing to check at this point.

Assumption 2: Random Sampling: The data is sourced from COVID-19 US state policy database (Raifman J, Nocka K, Jones D, Bor J, Lipson S, Jay J, and Chan P.).

The data is for all 50 states hence it is census i.e. we shall be running the statistical inference on entire population model.

Assumption 3: No perfect multi-collinearity: No need to explicitly check for perfect collinearity, because R will alert us if this rare condition happens.

In addition to the above, assess to what extent imperfect multicollinearity is affecting the inference by running Variance Inflation Factor. VIF > 10 indicates indicates imperfect multicollinearity is a problem.

The below analysis indicates that imperfect multicollinearity between independent variables is not a problem.

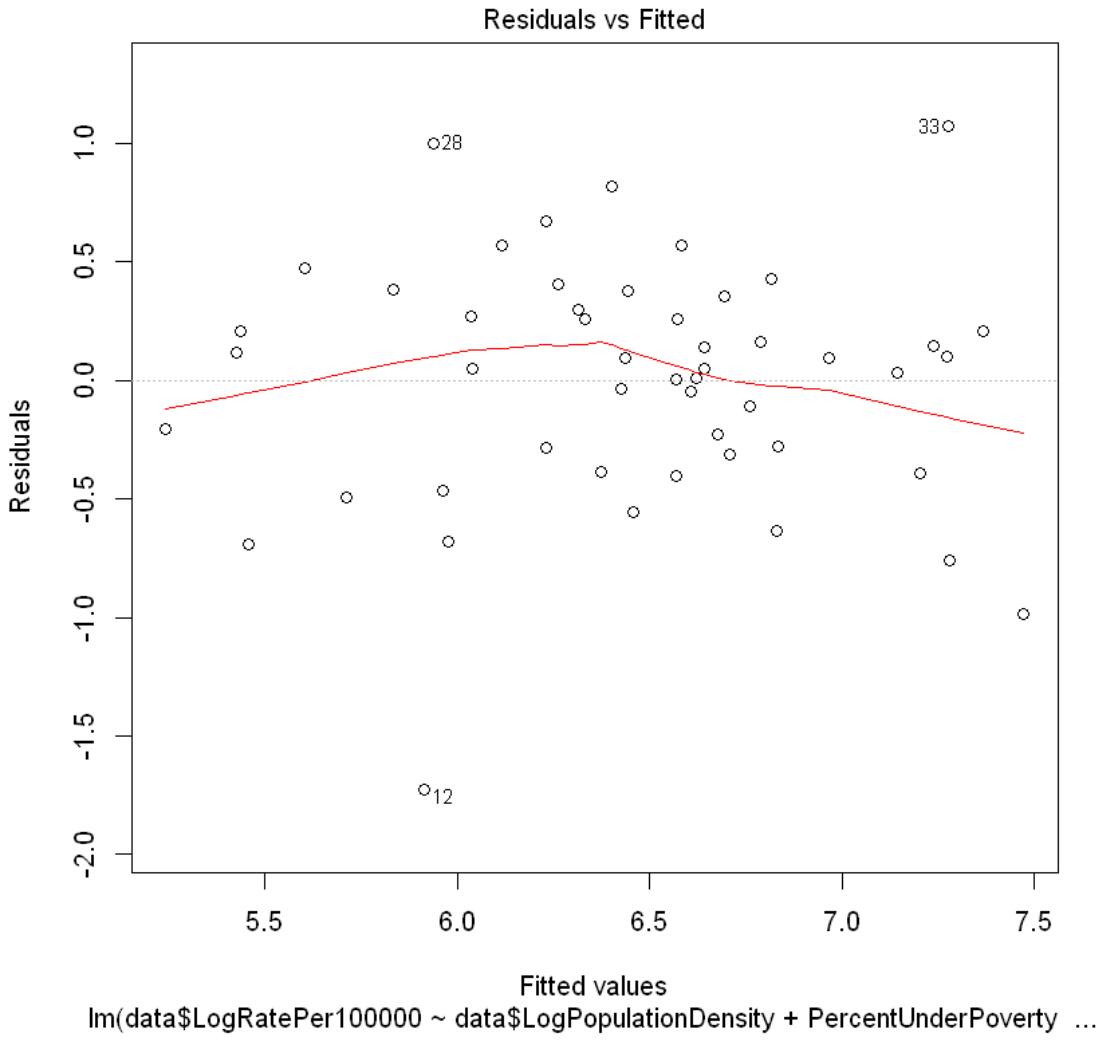
[30]:
```
vif(model3)
vif(model3) > 10
```

data\$LogPopulationDensity 1.06267078100033 **PercentUnderPoverty** 1.7870893472764
PercentAtRisk 1.81540798606853

data\$LogPopulationDensity FALSE **PercentUnderPoverty** FALSE **PercentAtRisk** FALSE

Assumption 4: Zero Conditional Mean: We start looking at the diagnostic plots:

[31]:
```
plot(model3, which = 1)
```



Covariance of Residuals with independent variables is close to 0. See analysis below. Hence the bias of the estimator will converge in probability to zero as $n \to \infty$, i.e. the OLS coefficients will be consistent.

[37]:
```
covPopulationDensity <- cov(model3$residuals,log(data$PopulationDensity))
paste("Covariance of Residual with PopulationDensity ", covPopulationDensity)
```

```
covPercentAtRisk <- cov(model3$residuals,data$PercentAtRisk)
paste("Covariance of Residual with PercentAtRisk ", covPercentAtRisk)

covPercentUnderPoverty <- cov(model3$residuals,data$PercentUnderPoverty)
paste("Covariance of Residual with PercentUnderPoverty ",␣
 ↪covPercentUnderPoverty)
```

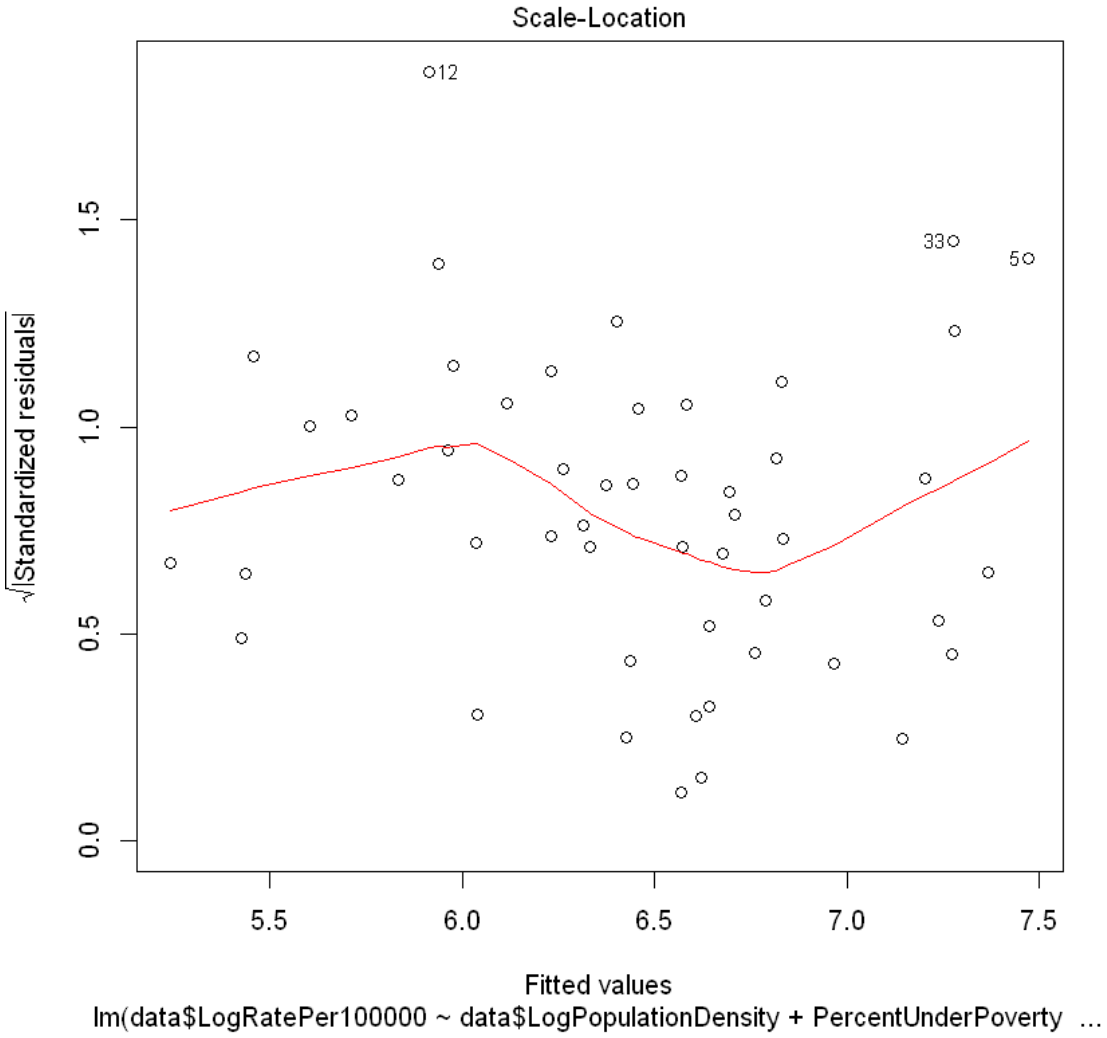
'Covariance of Residual with PopulationDensity 1.23123326301951e-17'

'Covariance of Residual with PercentAtRisk 1.42800489403815e-16'

'Covariance of Residual with PercentUnderPoverty -1.14436986141471e-16'

Assumption 5: Homoskedasticity: Residuals versus fitted values plot seems to indicate heteroskedasticity - the band seems to have uneven thickness. The scale location plot shall be used to assess this assumption, which indicates some heteroskedasticity but not significantly. we shall run Breusch–Pagan test to further validate this assumption.

[33]:
```
plot(model3, which = 3)
```



Running Breusch–Pagan test : Null hypothesis is for homoskedasticity

[34]:
```
bptest(model3)
```

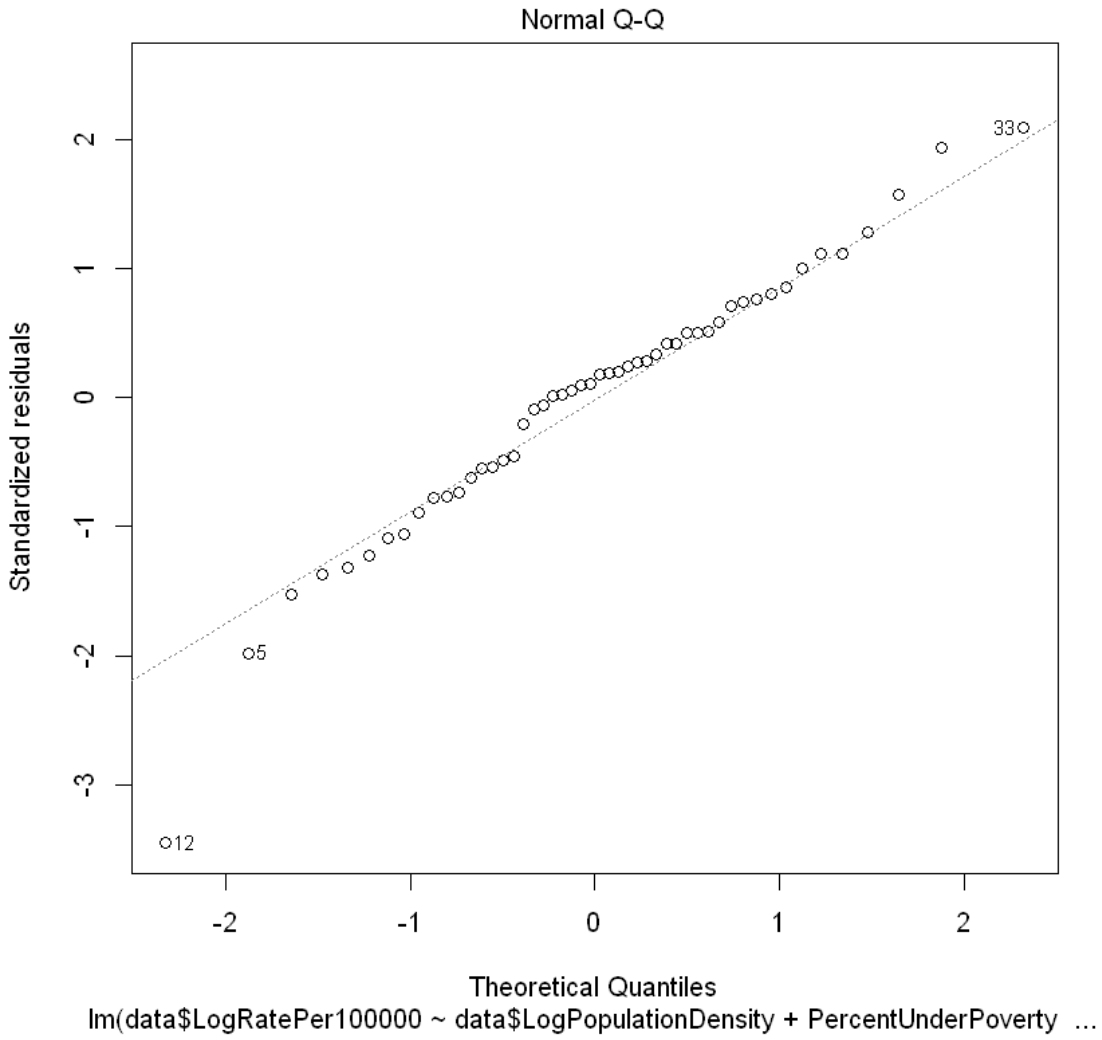
```
studentized Breusch-Pagan test

data:  model3
BP = 1.5692, df = 3, p-value = 0.6664
```

Since p-value > 0.05 hence we fail to reject null hypothesis. Despite meeting homoskedasticity condition we are using Huber-White's robust standard errors to be more conservative.

Assumption 6: Normality of Errors: To check normality of errors, we can look at the qqplot

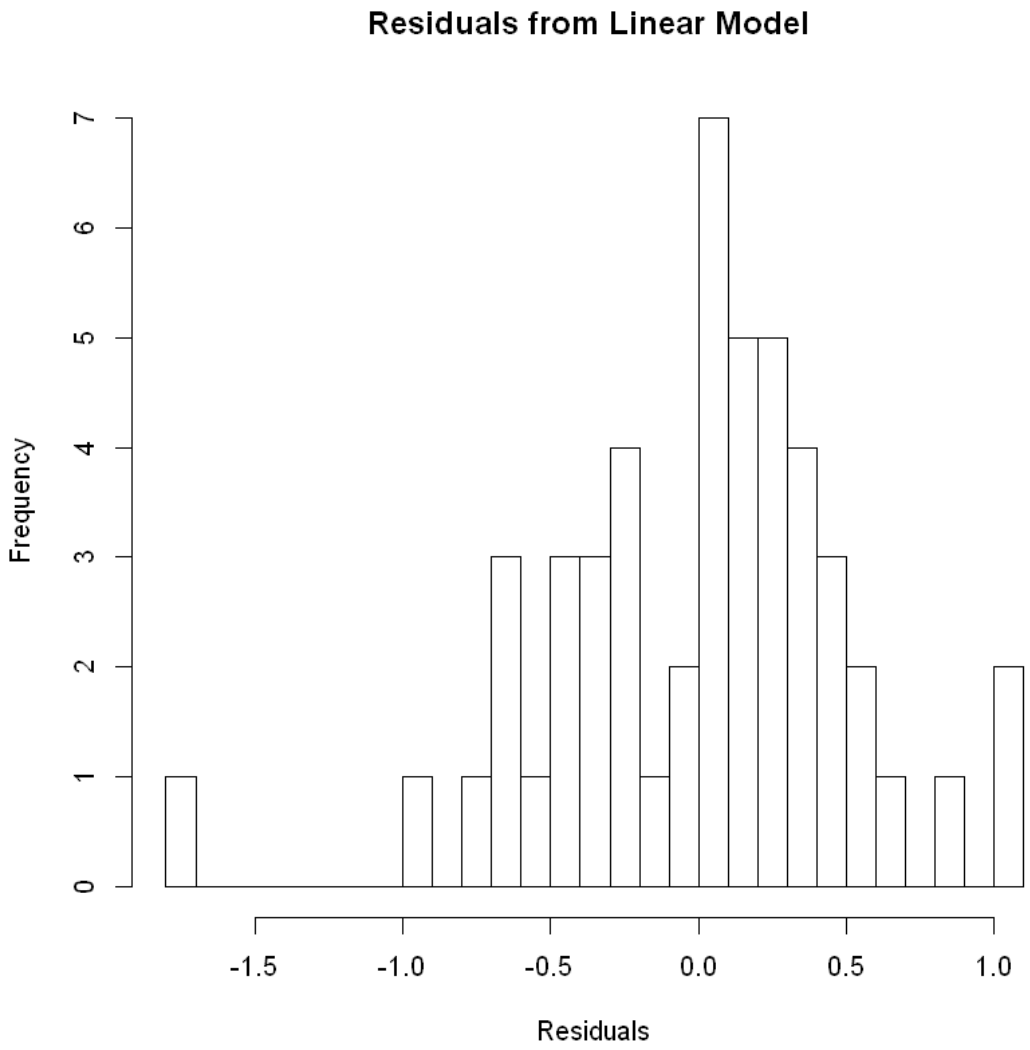
that's part of R's standard diagnostics

[35]:
```
plot(model3, which = 2)
```



Plotting histograms of residuals for visual verification

[36]:
```
hist(model3$residuals, breaks = 20, main = "Residuals from Linear Model ",␣
↪xlab="Residuals")
```



Both methods suggest we have a slight right skew but the data looks reasonably normally distributed.

3 Omitted variables

In this exercise, we are trying to model the population normalized Covid-19 infection rate per state. In our best model (V1) we are trying to account for :

1. Structural factors (i.e. the population density per state and percent population at risk) given that covid is a highly communicable infectious disease; and
2. Social factors (i.e. the percent of population under poverty) as these factors might affect the individual behavior whcih in turn might affect the overall infection rate

These features, however, are very coarse and present a very broad level picture. There are other

variables not present in our data set which, if captured, might present a better picture of the overall infection rate. In our detailed exploration of CML assumption of model3 above, we showcased that we are able to support all the assumptions of CLM reasonably well. However, we think that we might get some improvement (especially from the heteroskedasticity point of view) if some of the omitted variables are included in the analysis.

3.1 Omitted variable analysis Framework

We will consider one such feature (i.e. omitted variable) v in our analysis below to set up a framework for analysis.

$$\log(RatePer100000) = \beta_0 + \beta_1 \log(PopulationDensity) + \beta_2 PercentUnderPoverty + \beta_3 PercentAtRisk + \beta_4 v + u$$

$$v = \gamma_0 + \gamma_1 \log(PopulationDensity) + \gamma_2 PercentUnderPoverty + \gamma_3 PercentAtRisk + e$$

$$\log(RatePer100000) = (\beta_0 + \beta_4\gamma_0) + (\beta_1 + \beta_4\gamma_1) \log(PopulationDensity) + (\beta_2 + \beta_4\gamma_2) PercentAtRisk + (\beta_3 + \beta_4\gamma_3) Pe$$

3.2 Discussion of individual variables

Now we can consider one (probable) omitted variable at a time and see its effect on out computed OLS coefficients

1. **Ommited variable**: Level of adherence to social distancing rules:
    1. **Introduction**: This feature is hard to measure but it should be inversely correlated with ratePer100000 ($\beta_4 < 0$), inversely correlated with PopulationDensity ($\gamma_1 < 0$) as it will be hard to control big crowd behavior, inversly correlated with PercentUnderPoverty ($\gamma_2 < 0$) as it is hard for this socio economic section to remain indoors for long periods due to limited means and correlated with PercentAtRisk ($\gamma_3 > 0$) as fear will a major driving factor for this demographic to adhere to the norms
    2. **Affect on coefficients**:
        1. $\log(PopulationDensity)$ coefficient: As $\beta_4 < 0$ and $\gamma_1 < 0$, the coefficient is overestimated
        2. $PercentUnderPoverty$ cofficient: As $\beta_4 < 0$ and $\gamma_2 < 0$, the coefficient is overestimated
        3. $PercentAtRisk$ cofficient: As $\beta_4 < 0$ and $\gamma_3 > 0$, the coefficient is underestimated
2. **Ommited variable**: Percentage of local economy related to tourism:
    1. **Introduction**: This feature is easier to measure and it should be correlated with ratePer100000 ($\beta_4 > 0$), inversely correlated with PopulationDensity ($\gamma_1 < 0$) as movement should be from high population density to lower density, correlated with PercentUnderPoverty ($\gamma_2 > 0$) as most tourism centric economies are cyclical and attract poor population in the peak season and inversly correlated with PercentAtRisk ($\gamma_3 < 0$) as such places usually don't have most elderly population
    2. **Affect on coefficients**:
        1. $\log(PopulationDensity)$ coefficient: As $\beta_4 > 0$ and $\gamma_1 < 0$, the coefficient is underestimated

2. $PercentUnderPoverty$ cofficient: As $\beta_4 > 0$ and $\gamma_2 > 0$, the coefficient is overestimated
3. $PercentAtRisk$ cofficient: As $\beta_4 > 0$ and $\gamma_3 < 0$, the coefficient is underestimated

3. **Ommited variable**: Average number of out of state entrants:
    1. **Introduction**: This feature is easier to measure and it should be correlated with rate-Per100000 ($\beta_4 > 0$), correlated with PopulationDensity ($\gamma_1 > 0$) as usually more inflow of people is correlated with the population density, correlated with PercentUnderPoverty ($\gamma_2 > 0$) as more potential opportunities attract more people to do regular jobs at minimum wage and correlated with PercentAtRisk ($\gamma_3 > 0$) as our intuition is that the poor population will be at high risk
    2. **Affect on coefficients**:
        1. $\log(PopulationDensity)$ coefficient: As $\beta_4 > 0$ and $\gamma_1 > 0$, the coefficient is overestimated
        2. $PercentUnderPoverty$ cofficient: As $\beta_4 > 0$ and $\gamma_2 > 0$, the coefficient is overestimated
        3. $PercentAtRisk$ cofficient: As $\beta_4 > 0$ and $\gamma_3 > 0$, the coefficient is overestimated
4. **Ommited variable**: Major international ports of entry:
    1. **Introduction**: This feature is easier to measure and it should be correlated with rate-Per100000 ($\beta_4 > 0$), correlated with PopulationDensity ($\gamma_1 > 0$) as international hubs are known to be crowded places, inversly correlated with PercentUnderPoverty ($\gamma_2 < 0$) as with more opportunities, the average poverty level should be lower and inversly correlated with PercentAtRisk ($\gamma_3 < 0$) as such places usually have the best medical facilities
    2. **Affect on coefficients**:
        1. $\log(PopulationDensity)$ coefficient: As $\beta_4 > 0$ and $\gamma_1 > 0$, the coefficient is overestimated
        2. $PercentUnderPoverty$ cofficient: As $\beta_4 > 0$ and $\gamma_2 < 0$, the coefficient is underestimated
        3. $PercentAtRisk$ cofficient: As $\beta_4 > 0$ and $\gamma_3 < 0$, the coefficient is underestimated

4 Conclusion

In our exploration of statewide covid-19 data, we started with a common-sense model for estimating the normalized infection rate. We wanted to test the hypothesis that whether the public policy decisions such as shelter in place and mandatory masks have had an impact on the normalized infection rates over and above the commonsense socio-economic features. We were surprised to see that two most relevant public policy decisions (i.e. duration of shelter in place orders in states and the duration of mandatory mask policy in public places in the state) seem to have no statistically significant affect on normalized infection rate. Our interpretation is that most probably the adhearance to these policies is not high which is resulting in this strange result.

Also, surprisingly our model has a negative OLS coefficient for Percent population At Risk in the state which is counter-intuitive. Our interpretation of this fact is that in the first wave of infection, population density dominated the infection rate. In US, the high risk population is concentrated in the low density (Mid-west / South-West) region. Hence we are seeing this anomaly. As the situation evolves, we are looking at Florida and Georgia becoming the new epicenters for the pandemic. Hence we think that this negative OLS coefficient is an artifact of infection in flux. Once enough time has passed (e.g. $>$ 1 year), and we come back to this problem, we should see a positive coefficient for PercentAtRisk.

Overall, our final model, which takes into account $\log(PopulationDensity)$, $PercentUnderPoverty$ and $PercentAtRisk$ to predict normalized infection rate $\log(RatePer100000)$ seems to be doing a good job at the state level. The adjusted R^2 is around 0.5 and all the variables are highly statistically significant. Population density seems to have some impact on the infetion rate with 1% increase in population density resulting in around 0.3% increase in the infection rate. Percent under poverty seems to have a very high impact (1% increase resulting in $\sim$ 15% increase in normalized infection rate) which makes sense as this disease and the measures taken to reduce the spread take high toll on the lower strata of socio-economic ladder. We have already discussed our interpretation of (counter-intuitive) negative coefficient for the Percent At Risk feature on the normalized infection rate.