
DUE TO COVID-19 OUR OFFICES ARE CLOSED TO THE PUBLIC.
IF YOU HAVE ANY QUESTIONS ABOUT APPLYING FOR UNEMPLOYMENT INSURANCE BENEFITS GO ONLINE TO WWW.LABOR.NY.GOV OR CALL 1-888-209-8124
IF YOU NEED ASSISTANCE WITH EMPLOYMENT SERVICES, PLEASE VISIT www.JobZone.ny.gov FOR:
- JOB SEARCH
- RESUME WRITING
- CAREER EXPLORATION
- LIVE CHAT
OR CALL 718 613-3760
WE ARE YOUR DOL
NEW YORK STATE Department of Labor

Lab 3 : A Regression Study of COVID-19

John Woolley, Kevin Martin, Luis Bochner

1. Introduction

COVID-19 is a rapidly evolving disease, with the jury still out on its drivers and their respective (im)materiality. While the US is a COVID-19 flashpoint, its 50 states (plus DC) are experiencing substantially different outcomes with regards to COVID-19 cases. Our lab seeks to investigate the causes of variation in infection rates across these jurisdictions with an emphasis on public policy, while controlling for important features of demography such as population density and proportion of seniors (age 65+).

Specifically, we hypothesize a casual relationship between infection rates and government programs to maintain income among affected populations.

We believe *a priori* that less generous unemployment benefits lead to a higher relative utility of wages for working people, which may lead to one subjecting oneself to greater risk of infection. One potential implication of this model, if true, would be that even given a constant risk aversion towards Covid infection across the population, some individuals may choose to take this risk out of economic necessity, while others would not do so, and thus **prohibition of certain economic activities, as widely adopted across the United States, is an insufficient and incomplete policy response in the absence of sufficient subsidies to affected workers** in the face of these economic incentives. We also feel unemployment insurance is an actionable policy tool - increasing unemployment benefits is a "stroke of the pen" undertaking and materially easier to execute/enforce than alternatives such as statewide lockdowns.

Our research question is thus:

How much does the maximum dollar amount of unemployment insurance benefits contribute to the per capita rate of confirmed COVID cases in a jursidiction?

We concede that the costs of living vary widely across different jurisdictions, but lack sufficient data to further control benefits for this and other factors; a further discussion of omitted variable biases is present in Section 5. We also caveat that a "snapshot" dataset is not without challenges - time metrics regarding ongoing interventions (e.g. ongoing stay-at-home orders) had to be artificially bounded. Finally, our dependent variable of interest - weekly unemployment insurance payouts - were recorded by their max, while the mean would have been a more informative and representative statistic. That being said, the data still allowed us to conduct a detailed investigation and come up with an actionable conclusion.

On a qualitative note, we'd like to reiterate that we seek to explain, not predict, infection rates. Prediction is perhaps better suited for epidemiologists and/or non-linear techniques, given complex virus mutations and COVID's exponential/non-linear spread (e.g. "superspreader" events).

Transforms aside, we added no new variables to the dataset; all variables used belong to the COVID dataset produced by Majid Maki-Nayeri et al. References to the dataset and pictures used are on this report's final cell.

We find that there is a significant negative relationship between a jusdiction's maximum unemployment insurance amount and its per-capita covid cases. The investigation leading up to that conclusion follows below.

In [6]:
```
#install packages for analaysis (can comment out if you already have
 them)
#suppressMessages(install.packages('car'))
#suppressMessages(install.packages('lmtest'))
#suppressMessages(install.packages('sandwich'))
#suppressMessages(install.packages('e1071'))
#suppressMessages(install.packages('stargazer'))
#print("Packages were installed")

#load up packages into the notebook
library(readxl)
library(car)
library(lmtest)
library(sandwich)
library(e1071)
suppressMessages(library(stargazer)) #stargazer is noisy on startup
library(ggplot2)
print("Libraries were successfully attached")

#Read in excel (supress warnings because it doesn't like some of the
 formatting)
suppressWarnings(A <- read_excel("covid-19_dist0720 .xlsx", sheet="Co
vid-19"))
print("Excel data has been read-in")
```

```
[1] "Libraries were successfully attached"
[1] "Excel data has been read-in"
```

1.1 Load and Clean Data

We take a brief aside to address an obvious DQ issue - rows 3 and 4, "Arizona" and "arizona", contain identical values for all other columns except "Total Cases and Total Death". What to do?

We solve the issue through quick arithmetic. Diving Arizona's 2018 population by 100,000 and then multiplying it by its deaths/infections per 100,000 rate, we get at the respective totals we 'should' be looking at, i.e.

$\frac{7171646}{100000} \cdot 25.2 = 1807$ for total deaths and

$\frac{7171646}{100000} \cdot 1367.7 = 98087$ for total infections.

Neither A/arizona satisfies this condition, but summing both rows gives us 1,809 and 98,089 infections respectively. We proceed to combine the deaths and cases for the two Arizonas, and put the corrected values under the capital "A" Arizona for good grammar's sake. We then discard "arizona" to avoid doubble-counting.

In [7]:
```
#Combine "arizona" and "Arizona" deaths and infections
A[3,'Total Cases'] = A[3, 'Total Cases'] + A[4, 'Total Cases']
A[3,'Total Death'] = A[3, 'Total Death'] + A[4, 'Total Death']
A<-subset(A, State!="arizona")
```

2. Model Building and Specifications

Our outcome variable, confirmed COVID cases per 100,000 residents ("RatePer100000"), was decided on two grounds.

1. We chose rates over actuals in order to control for population, a necessary choice given the most populous state (California) has 68 times as many people as the least populous state (Wyoming).
2. We chose infections over deaths because it is the key variable in measuring COVID's spread and severity. Infection is a necessary condition for death, and hospitalizations coming from a higher infection rate can lead to sharp increases in mortality rates: As more strain is put on medical systems, suboptimal care standards and supply shortages affects' patients survival.

Having motivated our outcome variable, let’s zero in on our key explanatory variable to build our baseline model.

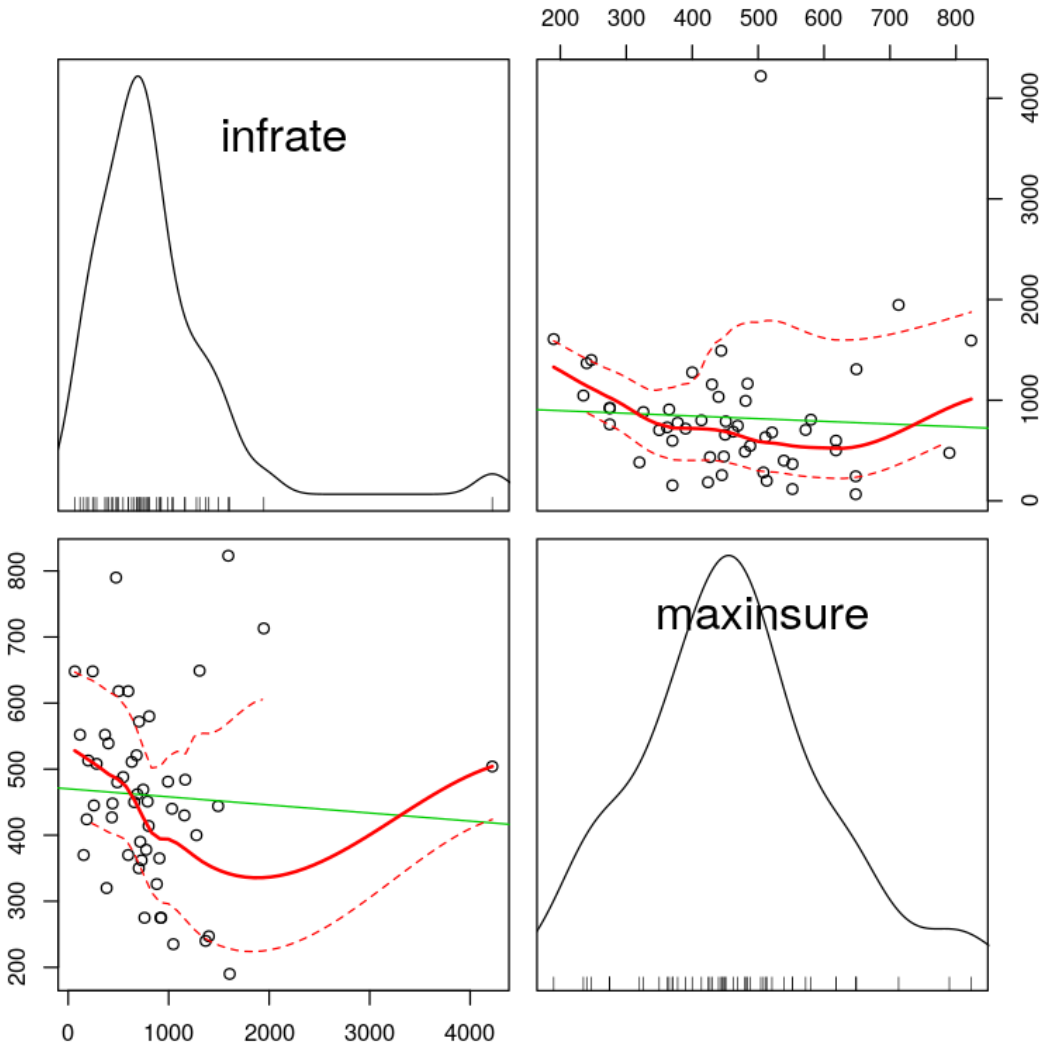
2.1 Baseline Model (V0)

In [8]:
```
#set up vectors for important regressors

# Confirmed covid cases per 100,000 residents of a given jurisdiction
infrate <- A$RatePer100000
# Maximum weekly unemployment rate (USD) for jurisdiction
maxinsure <- A$'Weekly unemployment insurance maximum amount (dollar
s)'
# Population density of jursdiction (people per square mile)
popdens <- A$'Population density per square miles'
```

2.1.1 Log-Log model

We introduced the pre-transformation dependent and independent variables in our introduction. We now consider whether to apply transformations to these covariates in light of EDA. We first scatterplot the variables as they are:

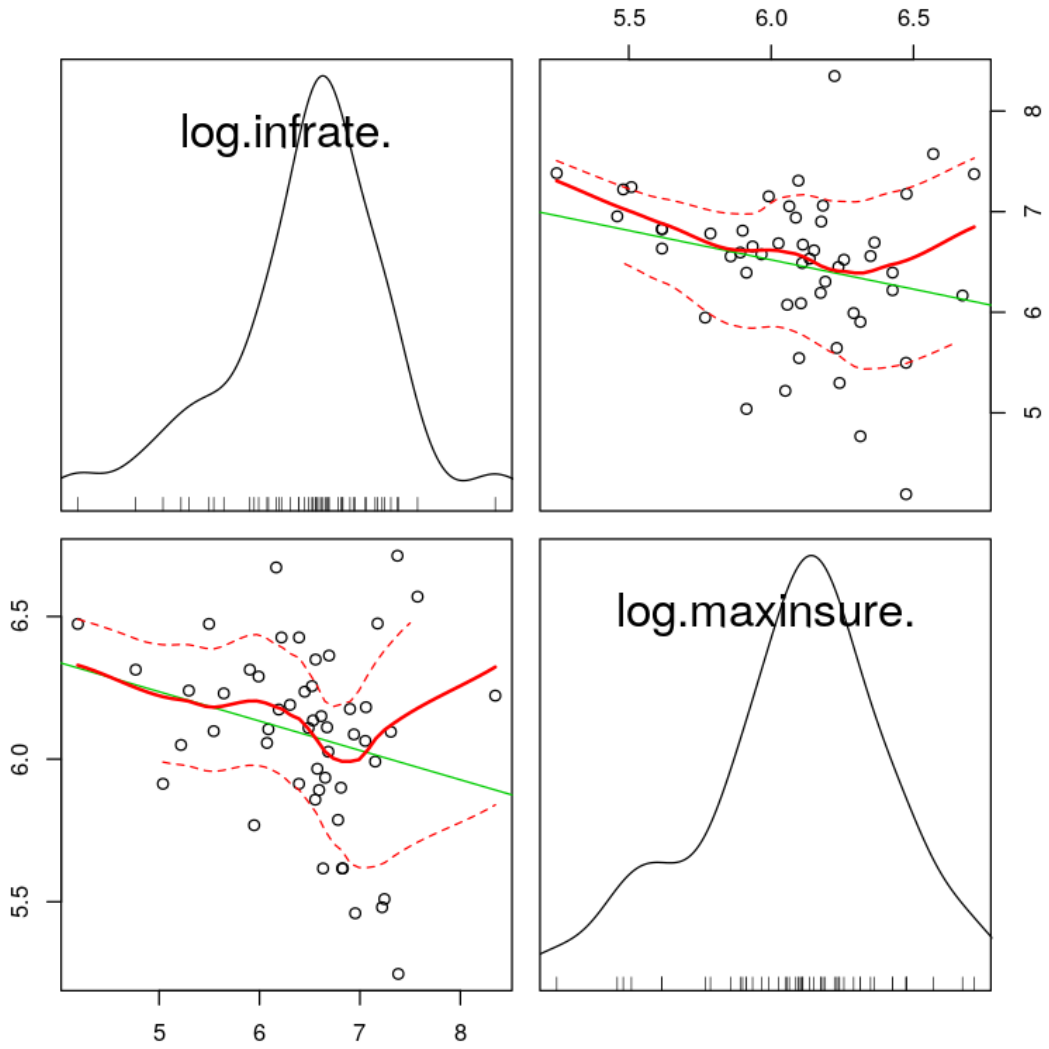
In [9]:
```
scatterplotMatrix(~infrate + maxinsure, data=NULL)
```



Observations like New York and DC could present challenges to our analysis. New York has double the rate of cases of New Jersey in second place, and the District of Columbia's population density is nearly an order-of-magnitude greater than other states due to its compact urban-only nature. New York is also affected by swaths of relatively very sparsely populated upstate regions whose population density and caseload contrast starkly with the beating heart of New York City.

A log-log model may help to reduce the leverage of these 'outliers' and lead to higher explanatory power. The relationship modeled by the log-log model roughly translates into a percent for percent elasticity of infection rates to maximum unemployment insurance payouts. We can also exponentiate the estimate of the log case rate to recover the estimated rate itself.

```
In [10]: scatterplotMatrix(~log(infrate) + log(maxinsure), data=NULL)
```



Given a tighter distribution as well as a more pronounced linear relationship, we choose to proceed with a log-log specification. There are qualitative benefits to this as well - much like with wages in labor economics, logging unemployment insurance has an intuitive takeaway of "an x% raise in benefits"; it also accounts for the fact that a dollar unit-increase means materially different things across states due to differences in purchasing power.

2.1.2 Baseline Model (V0)

We fit our model and see that our explanatory value, $\log(maxinsure)$ is significant at a 0.1 significance level under a 0.1 significance level under our baseline (but deliberately naive) model specification before further examining the model residuals for a brief discussion on CLM assumptions.

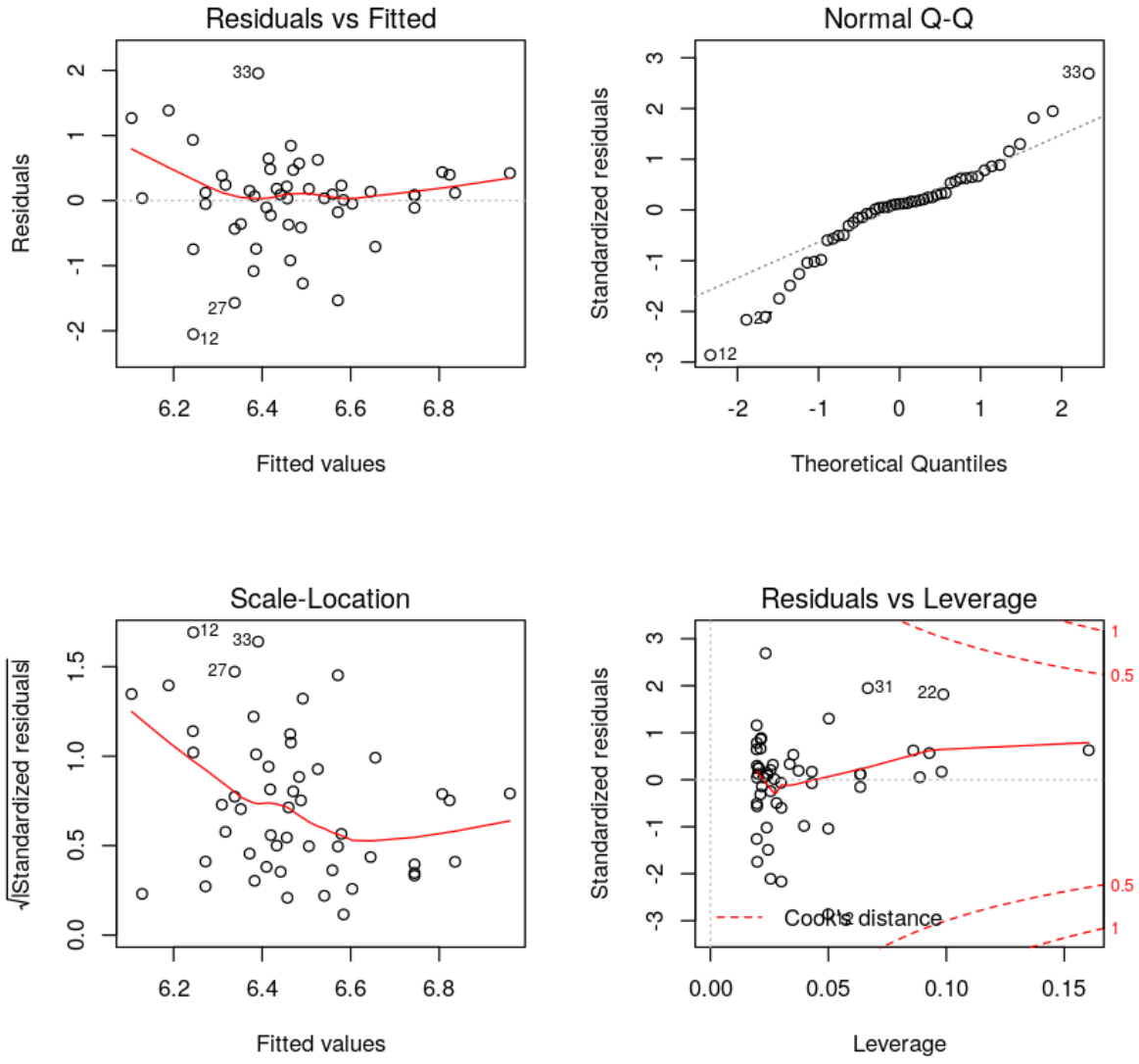
In [11]:
```
basemodel <- lm(log(infrate) ~ log(maxinsure))
summary(basemodel)
baseresid<-residuals(basemodel)
```

```
Call:
lm(formula = log(infrate) ~ log(maxinsure))

Residuals:
     Min       1Q   Median       3Q      Max 
-2.05328 -0.29369  0.08591  0.39059  1.95670 

Coefficients:
               Estimate Std. Error t value Pr(>|t|)    
(Intercept)     10.0192     2.0084   4.989 8.04e-06 ***
log(maxinsure)  -0.5831     0.3296  -1.769   0.0831 .  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7361 on 49 degrees of freedom
Multiple R-squared:  0.06002,   Adjusted R-squared:  0.04084 
F-statistic: 3.129 on 1 and 49 DF,  p-value: 0.08314
```

2.1.3 Baseline Model Residuals Analysis

1. We examine the **residuals plot** and see from the spline a deviation from zero conditional mean towards the extremities of the estimator (x) axis.
2. The **normal qq plot** reveals that the standardized residuals are definitely not normally distributed, with the tails exhibiting excess kurtosis.
3. The **scale-location** plot exhibits a red flag on heteroskedasticity, with the square root of the standardized residuals having a distinctly positve bias for lower values on the estimator axis.
4. We note that none of our standardized residuals exhibit excessive leverage (Cook's D < 1) and therefore no observations should be classified as outliers.

In [12]:
```
par(mfrow = c(2,2), oma = c(0,0,0,0)) #oma = outside margins
plot(basemodel)
```



The **Breusch-Pagan test** allows us to quantitatively test the degree of heteroskedasticity in the sample via hypothesis testing. In this test, the null hypothesis corresponds to homoskedasticity, which we reject.

In [13]:
```
bptest(basemodel)
```

```
        studentized Breusch-Pagan test

data:  basemodel
BP = 5.0263, df = 1, p-value = 0.02496
```

Motivated by our visual and quantitative findings, we use heteroskedasticity-robust standard errors and re-evaluate the significance of our estimator. Our heteroskedasticity-robust estimator is borderline not significant at the 0.1 level; the "naive" standard errors were in fact too small as they assumed homoskedasticity.

In [14]:
```
coeftest(basemodel, vcov = vcovHC)
(se.basemodel = sqrt(diag(vcovHC(basemodel)))) #robust SE
```

```
t test of coefficients:

               Estimate Std. Error t value  Pr(>|t|)
(Intercept)    10.01917    2.06355  4.8553 1.268e-05 ***
log(maxinsure) -0.58307    0.34847 -1.6732    0.1007
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

(Intercept): 2.063551997674 **log(maxinsure):** 0.348472270549144

However, we know there are key covariates we haven't considered yet that may address our concerns and will potentially lead to a more significant coefficient estimator for our key variable of interest, $\log(maxinsure)$.

2.2 Improved Model (V1)

After specifying and examining our baseline model, it is clear we are in need of more covariates.

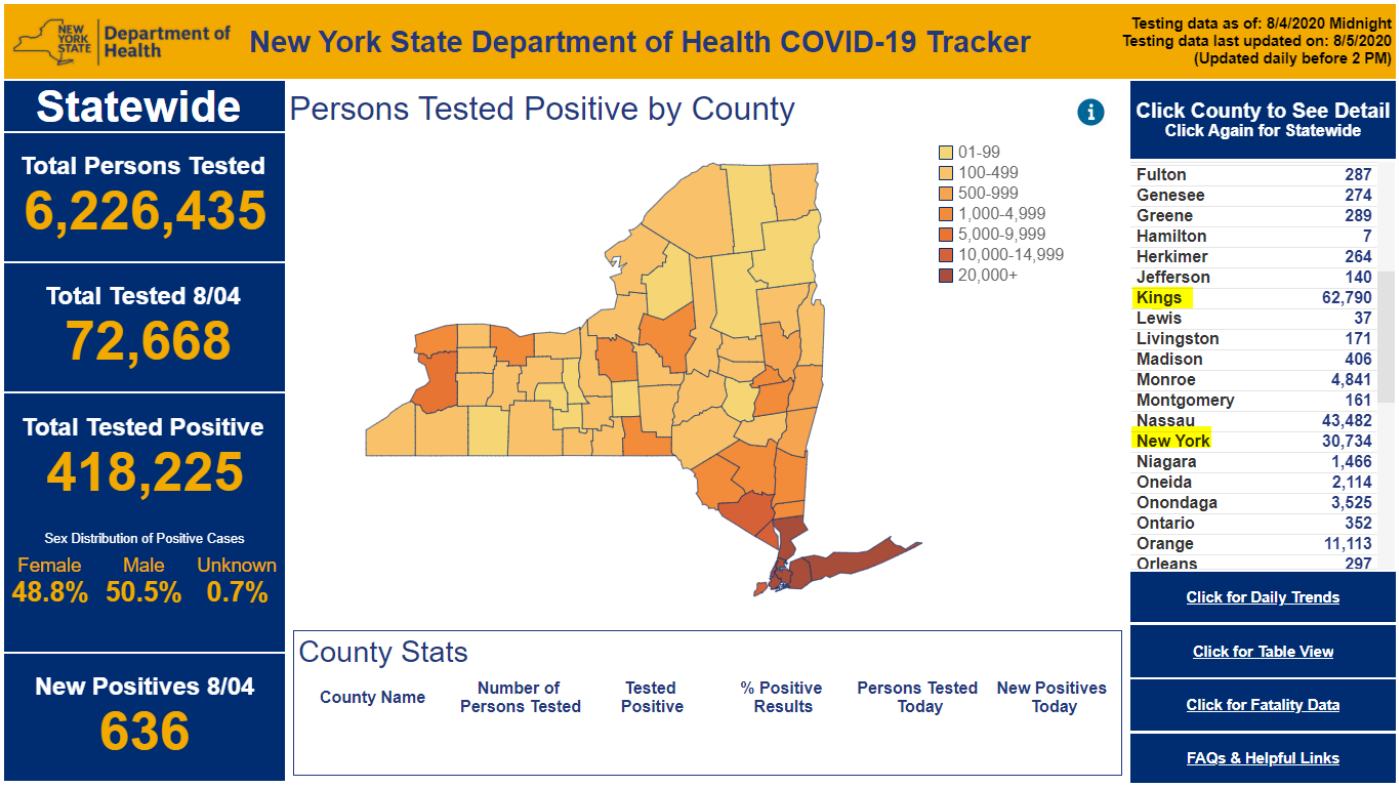
We now add two demographic variables that we believe are well-founded in theory and likely to be correlated with the residuals of the baseline model and not highly correlated with unemployment benefits, thereby offering additional explanatory power and increasing the significance of our slope estimate on our key variable of interest.

1. Log of Population Density
2. Proportion of Seniors (65+) in Population

2.2.1 Log of Population Density

We use New York as a case study as it is one of the most affected states, keeps good COVID data, and has (we hope) already experienced a full outbreak cycle.

The first, immediately noticeable takeaway is that the coronavirus thrives in crowded places, where airborne contagion is higher on average. See how the counties of New York (Manhattan) and Kings (Brooklyn) compare with the rest of the state. While not seen below, note that infection **rates** are materially larger as well:

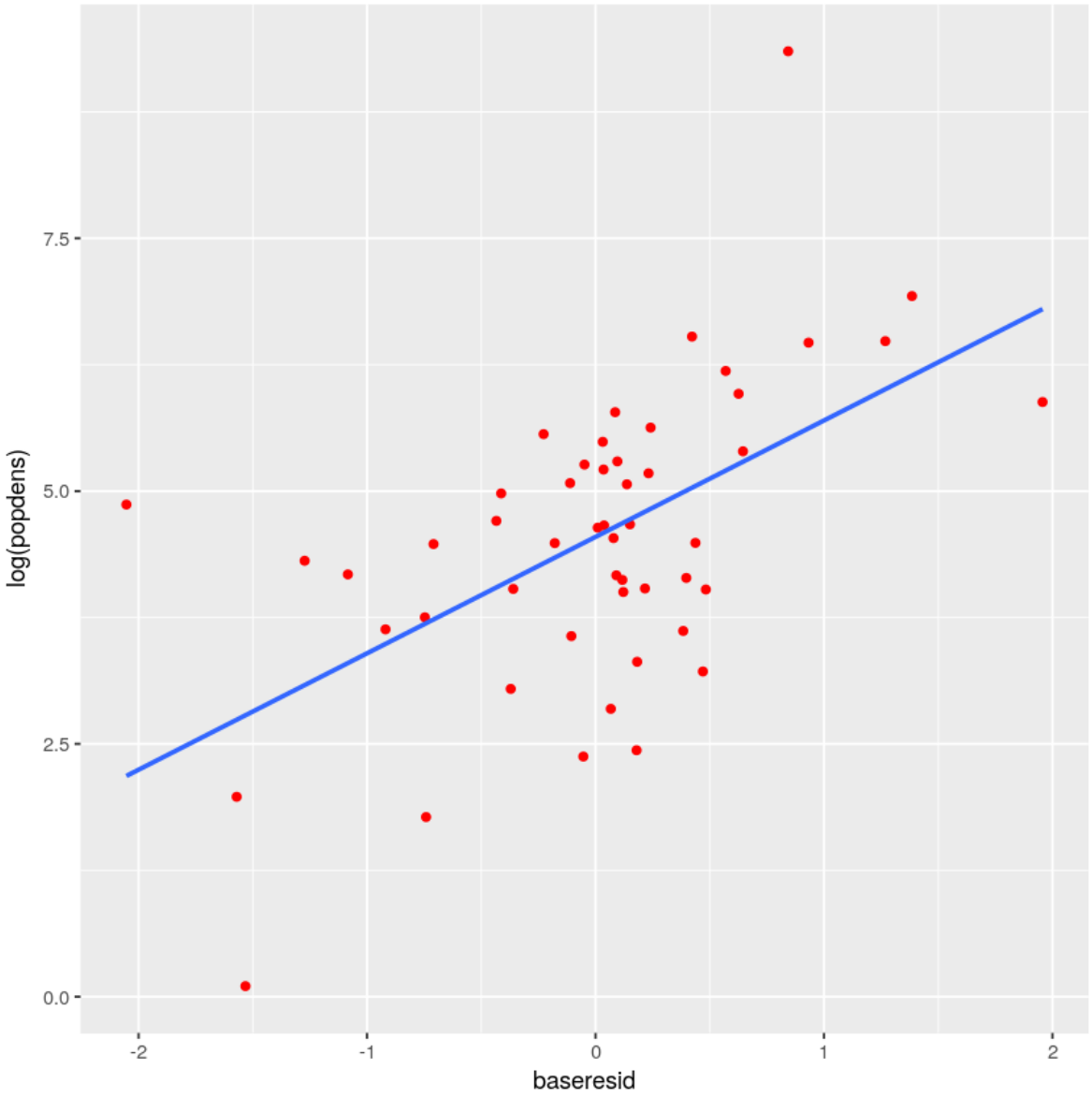


This trend holds for other, more recent flashpoints of the disease - they have all been in states with a number of population-dense cities, such as Florida or California.

We also note population density has a telling linear relationship with our residuals, and in the expected direction too - higher population densities should lead to higher infection rates.

We decide to log population density. This specification helps reduce the leverage of states like D.C, whose compact urban-only borders lead to a population density nearly an order of magnitude greater than the second densest state. We also notice that unit changes in population density would lead to a marginal coefficient given values are mostly in the hundreds.

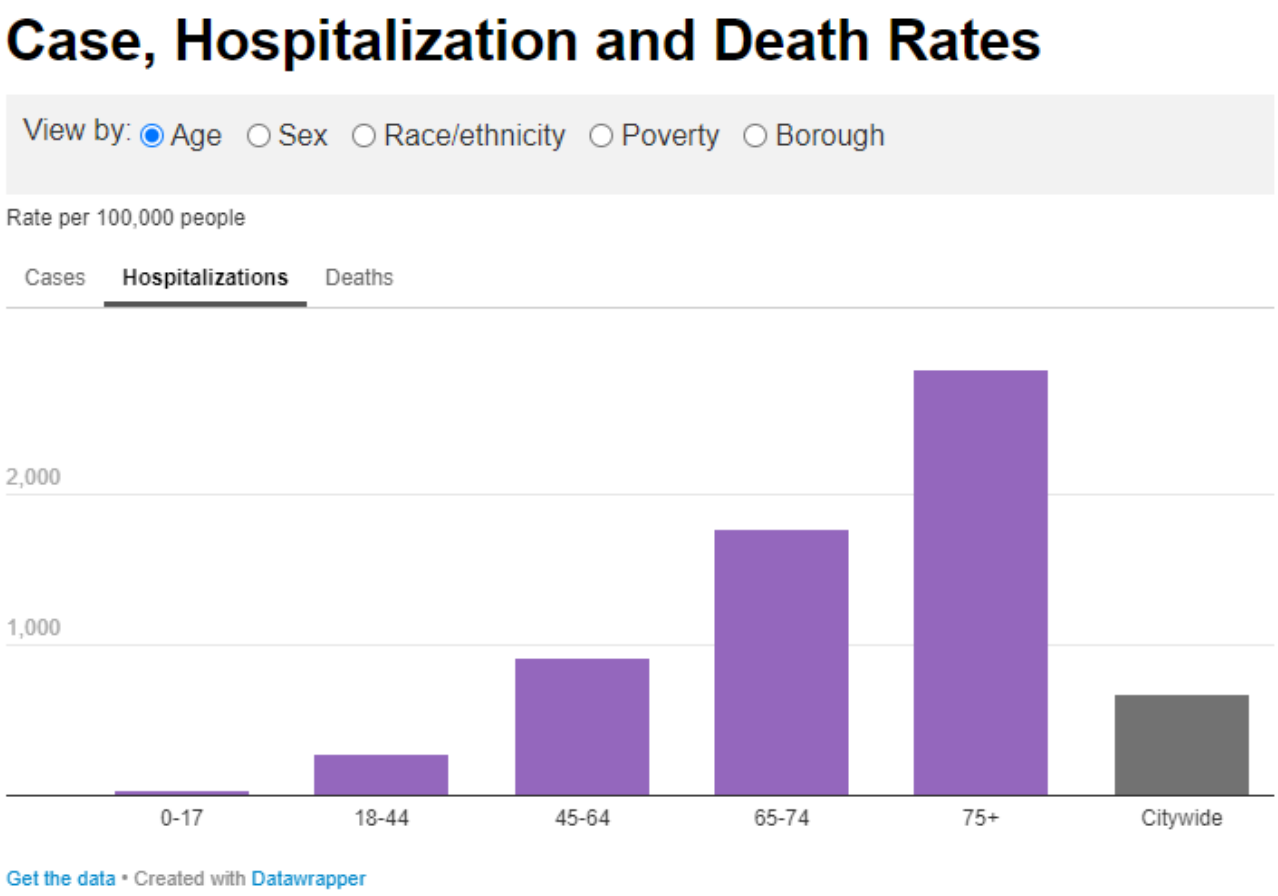
In [15]:
```
ggplot(aes(x = baseresid, y = log(popdens)), data=NULL) +
geom_point(colour = "red") +
geom_smooth(method = "lm", fill = NA)
```



2.2.2 Proportion of Elderly (65+) in Population

Drawing further inspiration from New York City demographic data, we recall how the elderly are especially vulnerable to both death and infection. The elderly are likely to have less robust immune responses than younger individuals, on average, making them more prone to infections in general. Because many elders are cognizant of this, it may also affect their behavior and make them more risk-averse than younger individuals. Specifically in the case of New York, the state government unfortunately mandated nursing homes continue to care for Covid-19 positive patients in inadequately-equipped facilities, and these nursing homes quickly became hotspots of contagion.

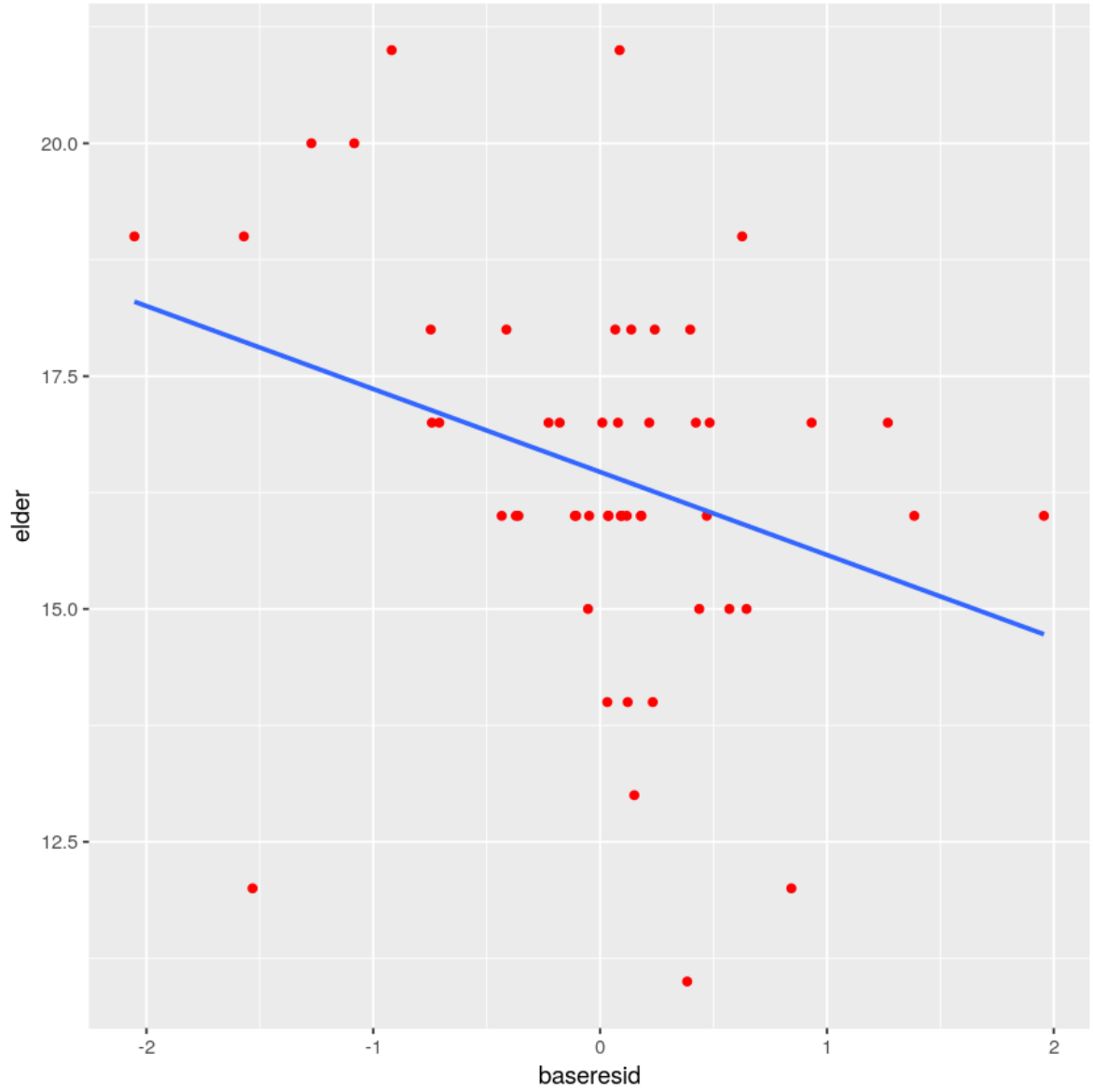
On the other hand, the elderly are more likely to be symptomatic and thus be captured in the data than younger, asymptomatic patients that never get tested.



We use the '65+' variable from the Maki-Nayeri et al. to operationalize this variable; it is defined as the percent of people aged 65 and above in a given jurisdiction and measures exactly what we want.

Visually, we see a linear dependence between the residuals of our base case model and our candidate regressor $elder$. Counterintuitively, the relationship is negative. It could be that elders' risk-averse behavior leads to a lower infection rate, or that our curret variable could unwittingly be a proxy variable for retirement: States that have larger proportions of retirees might expect lower contagion because less of their population needs to leave the house to work.

In [16]:
```
# Percent of population of juristiction who is elderly (over 65)
elder <- A$'65+'*100

ggplot(aes(x = baseresid, y = elder), data=NULL) +
geom_point(colour = "red") +
geom_smooth(method = "lm", fill = NA)
```



Because $elder$ is already a proportion, we see no need to apply a log transform.

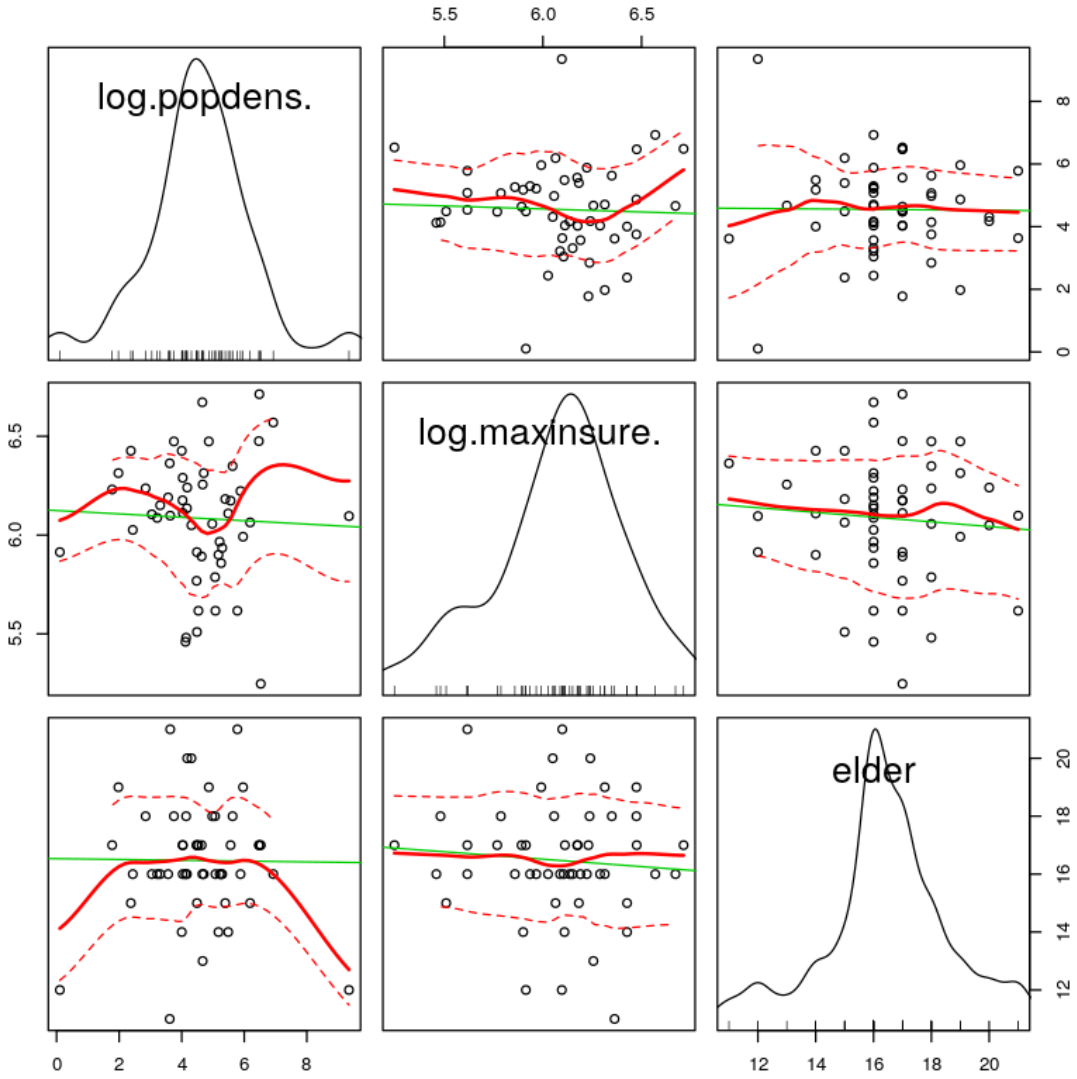
Indeed, this transform could well obscure our model as it would lead into a confusing "percent of percentages" interpretation.

2.2.3 Collinearity Check

As with any multivariate model, we run scatterplots to make sure there is no multicollinearity.

It looks like we have a sufficiently low degree of collinearity in our independent variables. While we see a few curvatures in the splines, we can see most of these are all associated with sparse data points and therefore likely statistical artefacts. Otherwise, we can barely see a discernible linear relationship.

In [17]:
```
scatterplotMatrix( ~ log(popdens) + log(maxinsure) + elder)
```



On a more directly quantitative note, we plot a correlation matrix and note that intercorrelations are quite weak among our key regressors.

In [18]:
```
cor(cbind(log(popdens),log(maxinsure),elder,deparse.level = 2))
```

A matrix: 3 × 3 of type dbl

	log(popden...	log(maxins...	elder
log(popden...	1.00000000	-0.04076164	-0.01001011
log(maxins...	-0.04076164	1.00000000	-0.07802489
elder	-0.01001011	-0.07802489	1.00000000

2.2.4 Fitting the Improved Model

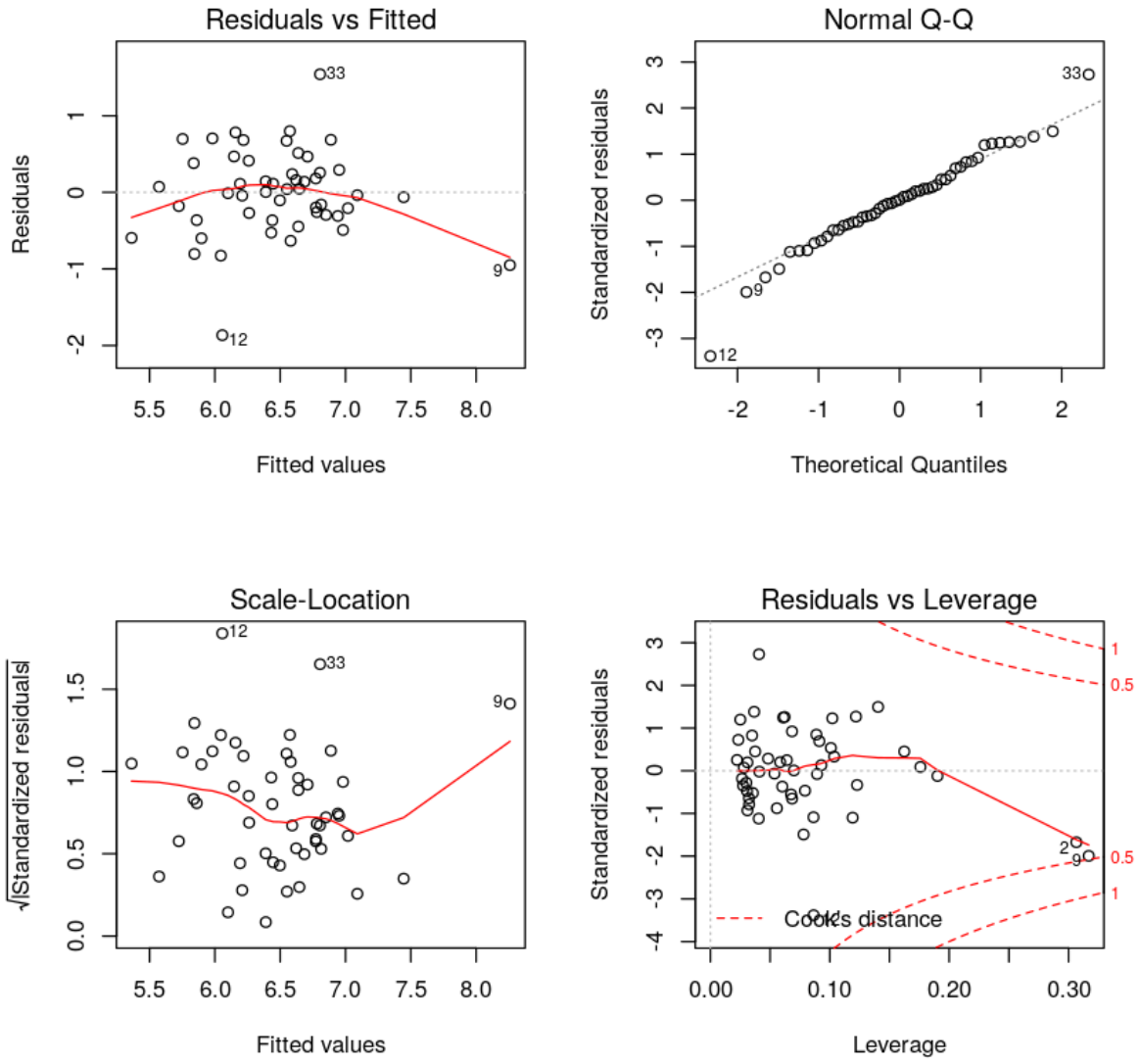
In [19]:
```
v1model <- lm(log(infrate) ~ log(popdens) + log(maxinsure) + elder)
summary(v1model)
```

```
Call:
lm(formula = log(infrate) ~ log(popdens) + log(maxinsure) + elder)

Residuals:
     Min       1Q   Median       3Q      Max 
-1.86565 -0.30182  0.00401  0.33690  1.54229 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)    10.57824    1.78381   5.930 3.43e-07 ***
log(popdens)    0.27283    0.05472   4.986 8.81e-06 ***
log(maxinsure) -0.58590    0.25942  -2.258  0.02860 *  
elder          -0.10822    0.03951  -2.739  0.00867 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5771 on 47 degrees of freedom
Multiple R-squared:  0.4459,    Adjusted R-squared:  0.4106 
F-statistic: 12.61 on 3 and 47 DF,  p-value: 3.582e-06
```

2.2.5 Improved Model Residuals Analysis

1. The **residuals plot** seems to have a conditional mean near zero for most values of the estimator, notwithstanding the residual on the most extreme estimated value. We can attribute this as a statistical artefact of sparse data.
2. The **normal qq plot** reveals that the standardized residuals are definitely not normally distributed, with the tails exhibiting excess kurtosis, although this deviation is relatively less than in the baseline model.
3. The **scale-location** plot exhibits a potential red flag on heteroskedasticity, with curvature that may be influenced by sparse observations around the extrema of the estimator. We will investigate further.
4. We note that none of our standardized residuals exhibit excessive leverage (Cook's D < 1) and none should be classified as outliers.

In [20]:
```
v1resid <- resid(v1model)
par(mfrow = c(2,2), oma = c(0,0,0,0)) #oma = outside margins
plot(v1model)
```



In [21]:
```
bptest(v1model)
```

```
        studentized Breusch-Pagan test

data:  v1model
BP = 4.1552, df = 3, p-value = 0.2452
```

Interestingly, our Breusch-Pagan test cannot lead us to reject the null hypothesis of homoskedasticity for our v1 model. However, we know from our plotting (i.e. the scale-location plot) that we cannot categorically demonstrate homoskedasticity.

Motivated by our findings, we use heteroskedasticity-robust standard errors and re-evaluate the significance of our estimator. Even with robust standard errors, we see a drastic improvement from our baseline model, with every coefficient estimator significant at the 5% level, and $log(popdens)$ even being significant at 1%.

In [22]:
```
coeftest(v1model, vcov = vcovHC)
(se.v1model = sqrt(diag(vcovHC(v1model)))) #robust SE
```

```
t test of coefficients:

                Estimate Std. Error t value  Pr(>|t|)
(Intercept)    10.578243   2.339183  4.5222 4.144e-05 ***
log(popdens)    0.272830   0.088390  3.0866   0.00339 **
log(maxinsure) -0.585904   0.280875 -2.0860   0.04243 *
elder          -0.108223   0.053244 -2.0326   0.04777 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

(Intercept): 2.33918326052318 **log(popdens):** 0.0883904483569202 **log(maxinsure):** 0.280874998068566 **elder:** 0.0532441606867676

To check for the effects of collinearity, we also display the VIF for our newly fitted model; the VIF statistics reconfirm what our correlation matrix has shown us. All factors are barely above 1 (the value which indicates no correlation whatsoever) and far away from 5 (the rule-of-thumb value indicating there is a collinearity issue).

In [23]:
```
vif(v1model)
#vif(coef(v1model), vcovHC(v1model), matrix(v1model) )
```

log(popdens): 1.00183994618242 **log(maxinsure):** 1.00787538805485 **elder:** 1.00630162557409

2.3 Inclusive Model (V2)

We now seek to build a "kitchen-sink" model that will err on the side of inclusion to check the robustness of our v1 model. While we believe that all of our inclusions are reasonable, the main objective here is to establish the robustness of coefficients across model specifications.

1. Poverty Rate (federal poverty definition) as of 2018
2. An indicator variable for the northeast corridor

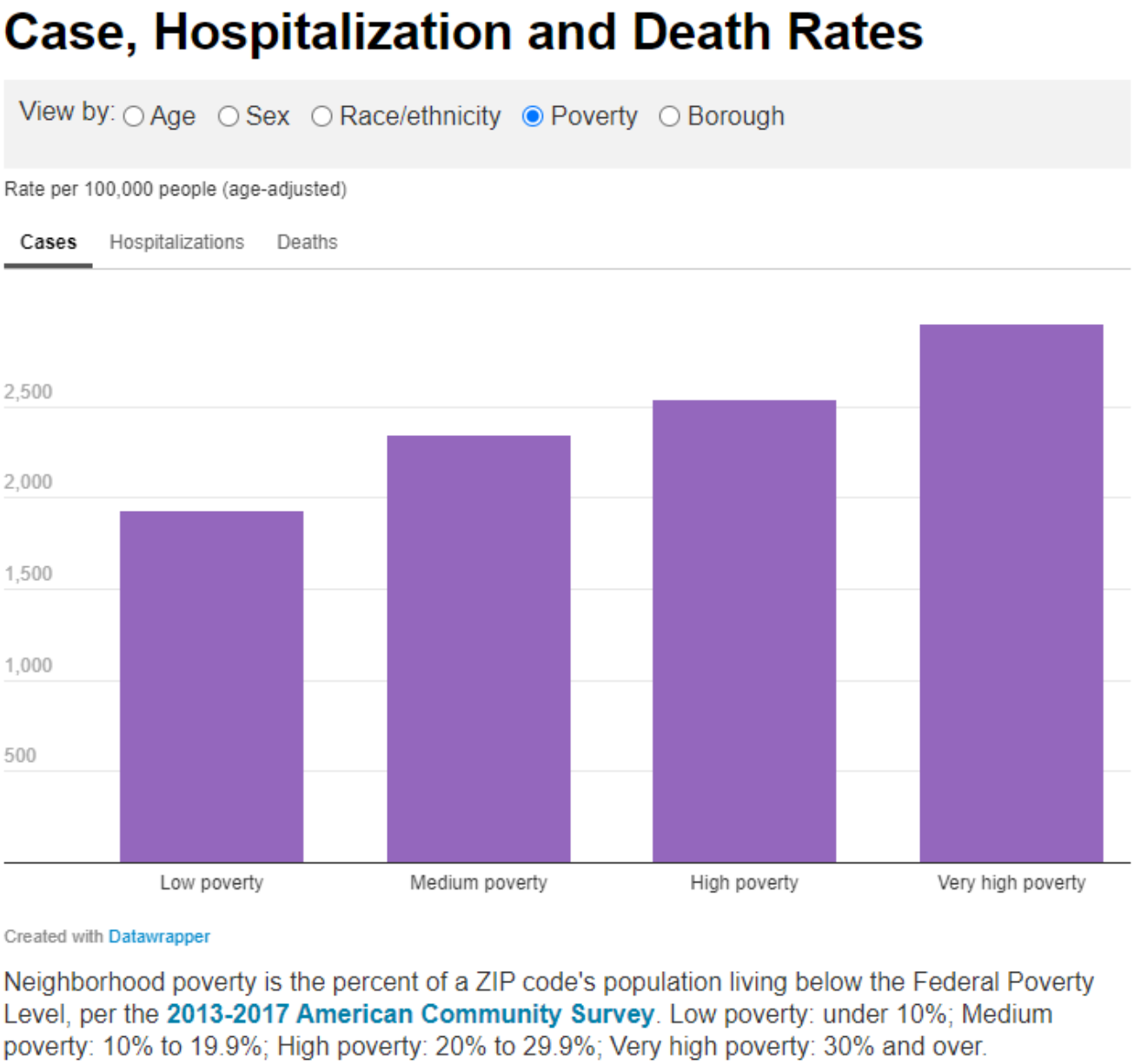
We also add a suite of variables to capture government interventions specific to covid-19 prevention:

1. Duration of shelter-in-place orders
2. Duration of non-essential business closures
3. Duration of mask-usage mandates
4. An indicator variable that is true where neither 3. nor 4. were implemented (duration==0)

One notable issue in this latter class of added covariates is that some of them exhibit what seems to be **reverse-causality**, i.e., it seems like some of these policies may be in place precisely because of high case rates, or a reasonable expectation of high case rates given other factors such as population density.

2.3.1 Poverty Rate

New York City data confirms a negative relationship between wealth and infection rates. This link may be due to several reasons: The need-to-work dimension we are currently exploring via our main dependent variable, poorer individuals' access to less and lower-quality healthcare, or just a lower immune system response due to poverty-related complications (e.g. inadequate nutrition, lack of sleep due to a longer commute).



We operationalize this via the "Percent living under the federal poverty line (2018)" variable from the Maki-Nayeri et al. dataset.

We note that this variable and maximum unemployment insurance may both capture a similar underlying "need to work" dynamic, however by directly measuring poverty levels we better capture the systemic link between poverty and health outcomes. (Conversely, a state may have

high/low unemployment benefits independent of their poverty levels.)

2.3.2 Northeast Corridor

We also code 'northeast_corridor', an indicator variable for the following states - NY, RI, MD, DC, PA, MA, NJ, DE, CT, VT,VI - which make up the "northeast corridor", also known as the "Acela corridor" for the eponymous train line shuttling tens of thousands of workers across these state lines every day (illustrated below). We know now that while patient zero in the US was in the West Coast, coronavirus outbreaks began in earnest in the northeast (NYC and the tri-state area) and most likely spread due to travel. Given the interconnectedness of this region, we model a dedicated indicator variable to control for this effect.

At the same time, we observe this indicator variable is likely to be correlated with population density and thus potentially interfere with its significance.

2.3.3 Covid-19 Pandemic-related Policy Interventions

Other covariates to include as in our "v2" model should definitely encompass government response - the extraordinary actions taken by governments should be expected to affect infection rates at least somewhat. We consider:

1. ‘shelter_days’, operationalized as the number of days between ‘End/relax stay at home/shelter in place’ and ‘Stay at home/ shelter in place’.

> When a start date but no end date is given, we impute ‘End’ values of 0 with July 2 2020, a.k.a when these data points were last updated. The ex-imputation dataset's mean and max respectively were 36 and 88 days, as opposed to 41 and 105 days post-imputation, a noticeable but not distorting change.

1. 'close_biz_days', a difference between ‘Began to reopen businesses statewide’ and ‘Closed non-essential businesses’.

> Note that 10 states made have no start dates for closing essential businesses, but do have end dates. We proxy these values by the average ‘State of emergency’ values. Businesses had to close before reopening began, and it is reasonable to assume closures as part of an emergency declaration. We assign a mean value given how tightly states of emergency are grouped, and in order to introduce as least variation as possible via imputation.

1. 'mask_days', a difference between ‘Mandate face mask use by employees in public-facing businesses’ and July 5th (the as-of date the variable was recorded).

> Note that 10 states had not enacted any mandate on face masks at the time of data collection; the value for this variable in these cases is 0.

1. 'covid_slackers' , an indicator variable based on scoring for how seriously (or not) states used government mandates to fight the virus. We classify as 1 states that did not enact a shelter-in-place order OR did not close non-essential businesses.

Finally, we refrain on using ‘State of emergency’ minus any particular date (e.g. the first recorded US case) as a covariate. All states declared an emergency within a roughly two-week period; we do not think that registers enough variation for a good regressor. In addition, the state of emergency is by itself a largely administrative measure; what truly may or may not move the needle is what is done with emergency resources/ powers/ etc.

In [24]:

```
#OBS: A WARNING ABOUT TIBBLE GETS GENERATED WHEN RUNNING THIS CELL. W
E VERIFIED
# ALL OUR LOGIC HELD FAST VIA A CSV EXPORT, AND BELIEVE THE WARNING D
OES NOT
# AFFECT OUR WORK.

A[ (A['Stay at home/ shelter in place'] != 0  &
    A['End/relax stay at home/shelter in place'] == 0),
    'End/relax stay at home/shelter in place' ] <- 44014

shelter_days <- (A['End/relax stay at home/shelter in place'] -
                 A['Stay at home/ shelter in place'] )

shelter_days <- unlist(shelter_days)

#Quick workaround to avoid type-wrangling: Calc mean(df$`State of eme
rgency`),
# put the date in excel, and insert it given our logic

A[ (A['Closed non-essential businesses'] == 0 &
    A['Began to reopen businesses statewide'] != 0),
    'Closed non-essential businesses'] <-  43900

close_biz_days <- unlist((A['Began to reopen businesses statewide'] -

                          A['Closed non-essential businesses'] ))

must_use_mask <- ifelse(
  A$`Mandate face mask use by employees in public-facing businesses`
!= 0, 1, 0)

# 44017 is the serial value for July 5, the as-of date for the mandat
e
# face mask variable
A['mask_days'] <-  44017 -
              A$`Mandate face mask use by employees in public-facing
 businesses`

A['mask_days'] <- ifelse(A$`mask_days` == 44017, 0, A$`mask_days`)

mask_days <- unlist(A['mask_days'])

A['northeast_corridor'] <- ifelse(A$State %in% c('Delaware','New Yor
k',
                            'Vermont', 'Maryland', 'Virginia', 'Maryla
nd',
                            'Pennsylvania', 'New Jersey', 'Connecticu
t',
                            'Massachussets', 'Rhode Island',
                            'District of Columbia') , 1, 0)

northeast_corridor <- unlist(A['northeast_corridor'])


A['covid_slackers'] = ifelse( ((A$`Closed non-essential businesses` =
= 0) |
```

```
                      (A$`Stay at home/ shelter in place` == 0)),1,0
)

covid_slackers <- unlist(A['covid_slackers'])
```

```
Warning message:
“The `i` argument of ``[<-`()` can't be a matrix as of tibble 3.0.0.
Convert to a vector.
This warning is displayed once every 8 hours.
Call `lifecycle::last_warnings()` to see where this warning was genera
ted.”
```

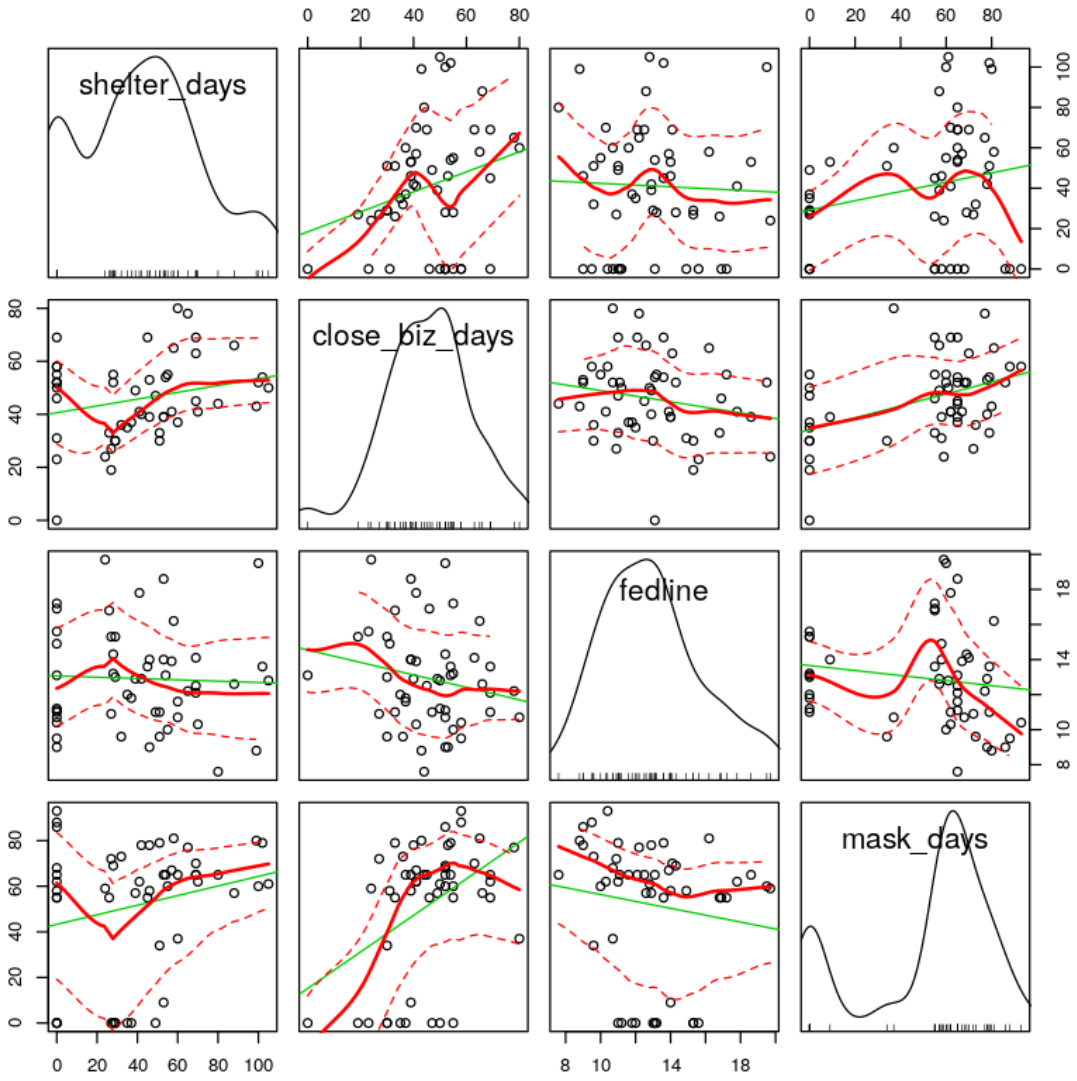
2.3.4 Collinearity Check

As with our other models, we run a collinearity check. Since scatterplots will get crowded, we first run scatterplots on the new covariates with themselves - note indicator variables are excluded as they aren't visually informative. We also run a correlation matrix to check for collinearity more directly:

In [25]:
```
fedline <- A$`Percent living under the federal poverty line (2018)`

A['fedline'] <- fedline
A['shelter_days'] <- shelter_days
A['close_biz_days'] <- close_biz_days

scatterplotMatrix(A[ , c("shelter_days", "close_biz_days", "fedline",

                         'mask_days')])


cor(cbind(fedline,shelter_days,close_biz_days,deparse.level = 2))
```

A matrix: 3 × 3 of type dbl

	fedline	shelter_da...	close_biz_...
fedline	1.00000000	-0.04038946	-0.1935502
shelter_da...	-0.04038946	1.00000000	0.2517885
close_biz_...	-0.19355017	0.25178848	1.0000000

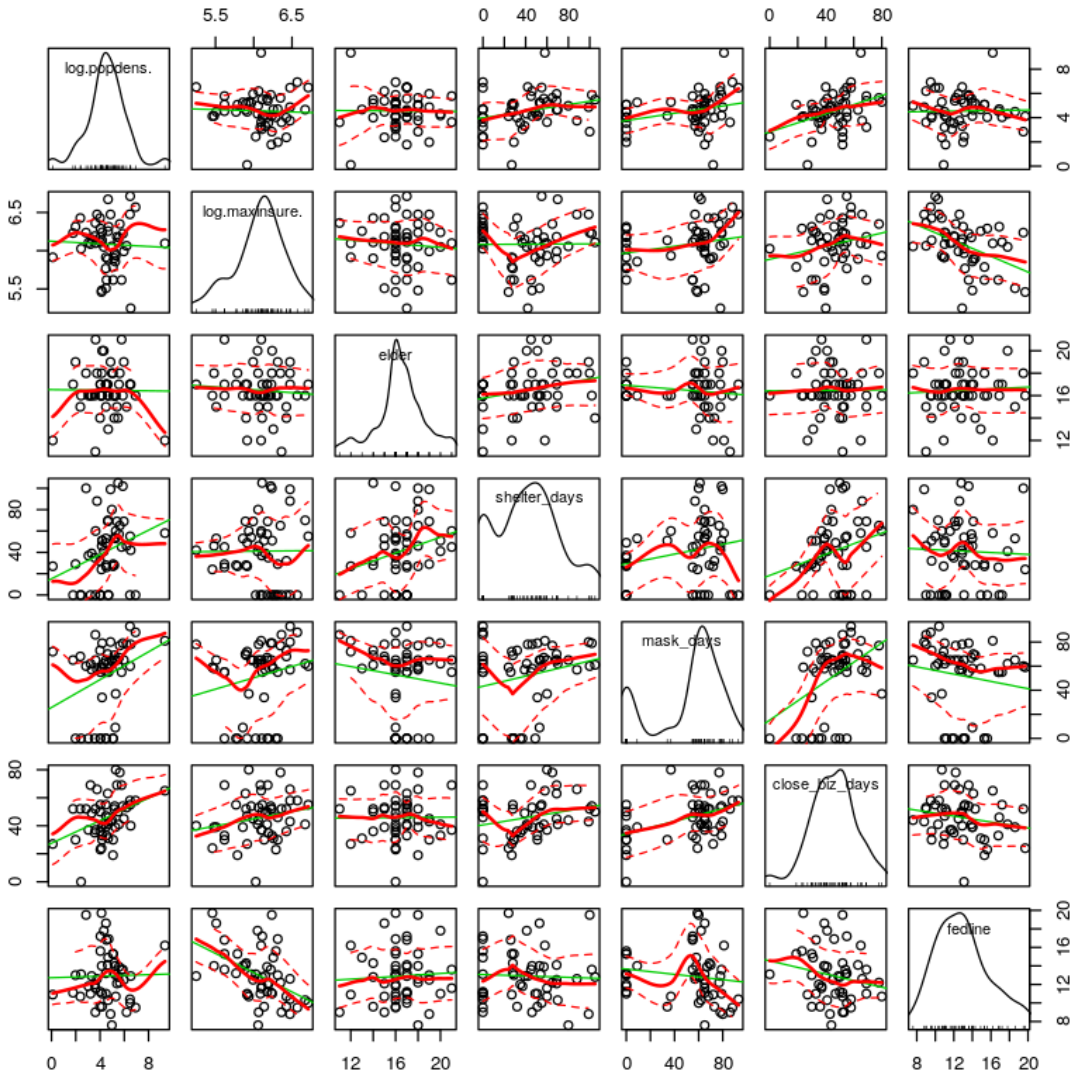


Collinearity is clear among our two time metrics, a not entirely surprising finding - one would expect states with short/long shelter-in-place orders to be similarly lax/punitive regarding businesses. (Our imputed end date may have contributed to this as well.)

These two variables' linear relationship with the poverty line variable is present, but comparatively subdued.

We are relieved to find no particular red flags. Now, to look at these new covariates with our existing v1 regressors:

In [26]:
```
scatterplotMatrix( ~ log(popdens) + log(maxinsure) + elder + shelter_
days +
                        mask_days + close_biz_days  + fedline, data=NUL
L)
```



Some linear relationships are practically inevitable; one would expect impoverished states to offer less unemployment insurance. Others make intuitive sense: It seems governors of dense and/or old states were also the ones taking lockdowns most seriously. We also see that poorer states enacted shorter lockdowns, which is a nod to the need-to-work dynamic we are measuring via our unemployment insurance variable.

Reassuringly, unemployment insurance seems uncorrelated for all other variables except for the poverty line.

Finally, we also run a correlation matrix for a more quantitative check on collinearity. Again, we find some of the stronger correlations to be intuitive: Northeast corridor states are economically developed and densely populated, so a relatively high correlation between our indicator variable and our density and unemployment insurance variables is expected. The same applies to the strong negative correlation between our covid_slackers variable and the amount of days since mandatory mask usage: states that are reluctant to implement any one policy to fight coronavirus are likely reluctant to implement any of them.

In [27]:
```
cor(cbind(log(popdens),log(maxinsure),elder, shelter_days, mask_days,
 close_biz_days,
          fedline, covid_slackers, northeast_corridor, deparse.level
= 2))
```

A matrix: 9 × 9 of type dbl

	log(popden...	log(maxins...	elder	shelter_da...	mask_days
log(popden...	1.00000000	-0.040761638	-0.010010109	0.279136577	0.29551633
log(maxins...	-0.04076164	1.000000000	-0.078024894	0.007025362	0.20380228
elder	-0.01001011	-0.078024894	1.000000000	0.252819524	-0.12206885
shelter_da...	0.27913658	0.007025362	0.252819524	1.000000000	0.21962663
mask_days	0.29551633	0.203802282	-0.122068845	0.219626627	1.00000000
close_biz_...	0.38883833	0.210379034	0.008629577	0.251788477	0.42514463
fedline	0.02162240	-0.462042147	0.056118782	-0.040389465	-0.14709170
covid_slac...	-0.19623262	0.340846937	-0.239866635	-0.757727890	0.01074245
northeast_...	0.56570220	0.053933279	0.031149060	0.133864047	0.40321074

2.3.5 Fitting the Inclusive Model

We have what we need to run v2. The construction and run of our inclusive model follows below:

In [28]:

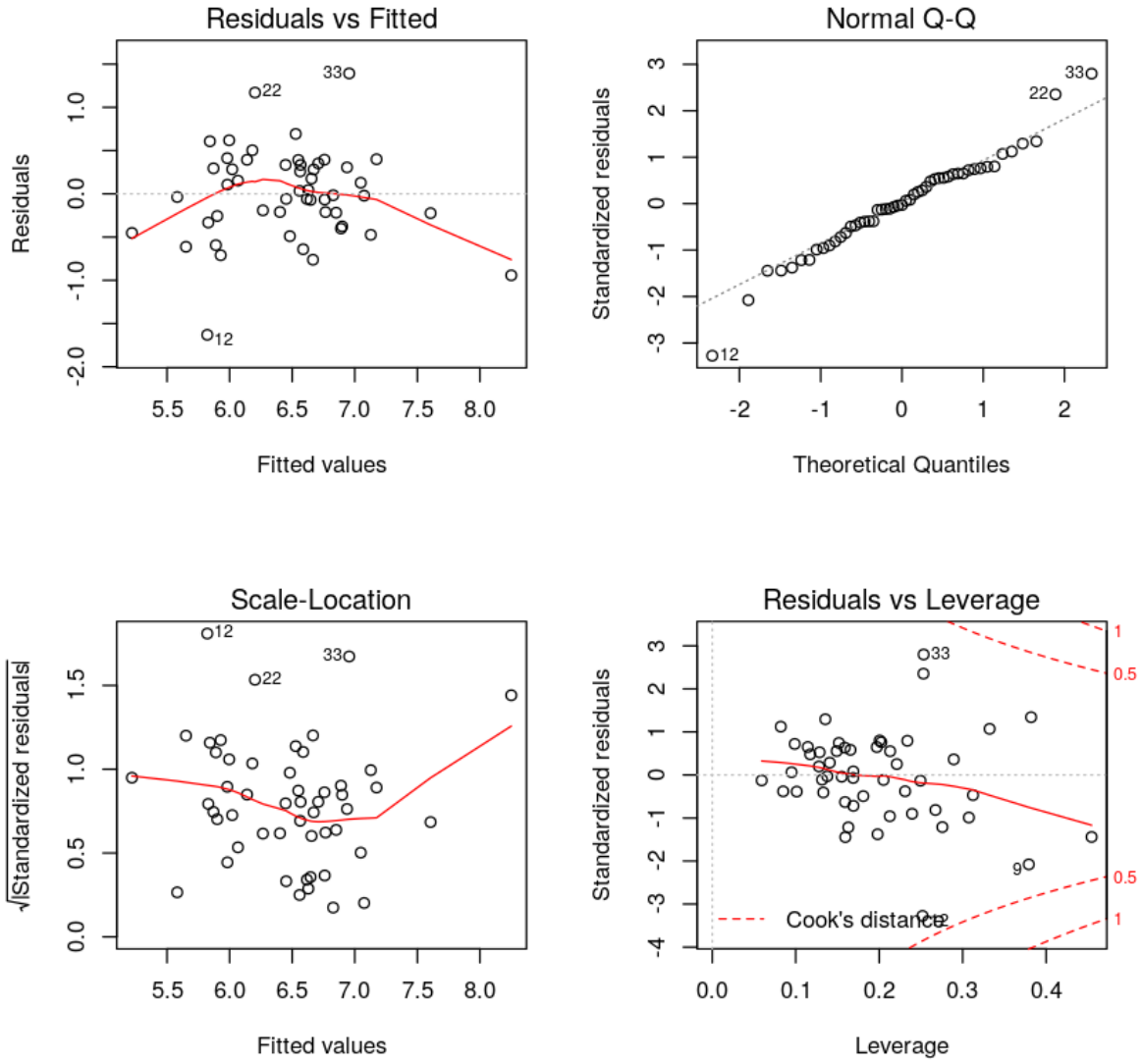
```
shelter_days <- unlist(shelter_days)
mask_days <- unlist(mask_days)
close_biz_days <- unlist(close_biz_days)
covid_slackers <- unlist(covid_slackers)

v2model <- lm(log(infrate) ~ log(popdens) + log(maxinsure) + elder +
              close_biz_days + shelter_days + mask_days + covid_slack
ers +
              fedline + northeast_corridor)

#note heteroskedasticity robust standard errors are not used at this
 point
summary(v2model)
v2resids <- residuals(v2model)
```

```
Call:
lm(formula = log(infrate) ~ log(popdens) + log(maxinsure) + elder +
    close_biz_days + shelter_days + mask_days + covid_slackers +
    fedline + northeast_corridor)

Residuals:
     Min       1Q   Median       3Q      Max
-1.63055 -0.29436 -0.01626  0.33457  1.39235

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)       11.543981   2.338787   4.936 1.38e-05 ***
log(popdens)       0.234503   0.072080   3.253  0.00229 **
log(maxinsure)    -0.835245   0.343373  -2.432  0.01945 *
elder             -0.098894   0.041816  -2.365  0.02284 *
close_biz_days     0.005149   0.006385   0.806  0.42472
shelter_days       0.002318   0.004988   0.465  0.64459
mask_days         -0.001561   0.003393  -0.460  0.64784
covid_slackers     0.537701   0.358687   1.499  0.14151
fedline            0.010991   0.033896   0.324  0.74739
northeast_corridor 0.271331   0.268582   1.010  0.31831
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5757 on 41 degrees of freedom
Multiple R-squared:  0.519,     Adjusted R-squared:  0.4135
F-statistic: 4.916 on 9 and 41 DF,  p-value: 0.0001723
```

2.3.6 Inclusive Model Residuals Analysis

1. The **residuals plot** seems to have a conditional mean near zero for most values of the estimator, notwithstanding the residual on the most extreme estimated value. We can attribute this latter as a statistical artefact of sparse data. We note the curvature is now less extreme than in the simpler model (v1).
2. The **normal qq plot** reveals that the standardized residuals are definitely not normally distributed, with the tails exhibiting excess kurtosis. There does not seem to be much improvement versus the last model examined, indeed the results here look slightly less normally distributed.
3. The **scale-location** plot exhibits a potential red flag on heteroskedasticity, with curvature that may be influenced by sparse observations around the extrema of the estimator. We will investigate further as in previous sections.
4. We note that none of our standardized residuals exhibit excessive influence (Cook's D < 1) and none should be classified as outliers.

In [29]:
```
par(mfrow = c(2,2), oma = c(0,0,0,0)) #oma = outside margins
plot(v2model)
```



In [30]:
```
bptest(v2model)
```

```
        studentized Breusch-Pagan test

data:  v2model
BP = 13.598, df = 9, p-value = 0.1374
```

Once again the Breusch-Pagan test contradicts our visual intuition by failing to reject homoskedasticity, although it does so at a lower level of confidence versus v1. Regardless, we proceed with heteroskedasticity-robust standard errors for the same reason as before and re-evaluate the significance of our coefficient estimators.

While we do not find that any of the additional covariates are significant, we find that most of our covariates from the simpler model still are a significant at the 10% level and nearly so at 5%, but $elder$'s p-value is a couple of percentage points shy of 10% significance. With 2/3 of our original regressors at an acceptable level of statistical significance, we see our inclusive model as confirmation that our v1 model is reasonably robust.

In [31]:
```
coeftest(v2model, vcov = vcovHC)
(se.v2model = sqrt(diag(vcovHC(v2model))))
```

```
t test of coefficients:

                    Estimate Std. Error t value Pr(>|t|)
(Intercept)       11.5439814  3.1274645  3.6912 0.000651 ***
log(popdens)       0.2345033  0.1214029  1.9316 0.060339 .
log(maxinsure)    -0.8352447  0.4334827 -1.9268 0.060952 .
elder             -0.0988943  0.0618785 -1.5982 0.117677
close_biz_days     0.0051485  0.0077208  0.6668 0.508616
shelter_days       0.0023179  0.0089421  0.2592 0.796764
mask_days         -0.0015615  0.0036302 -0.4301 0.669352
covid_slackers     0.5377010  0.5216276  1.0308 0.308671
fedline            0.0109909  0.0501275  0.2193 0.827536
northeast_corridor 0.2713305  0.4351143  0.6236 0.536355
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

(Intercept): 3.12746453372754 **log(popdens):** 0.121402921489792 **log(maxinsure):** 0.433482684398309 **elder:** 0.0618784660068494 **close_biz_days:** 0.00772084813644509 **shelter_days:** 0.0089420688898626 **mask_days:** 0.00363024921844112 **covid_slackers:** 0.521627577387504 **fedline:** 0.050127533712175 **northeast_corridor:** 0.43511429652405

To check for the effects of collinearity, we also display the VIF for our newly fitted model. We see an uptick across the board for our regresors' VIFs , which is somewhat expected due to increased intercorrelations, but almost all values are still less than two. Two regressors, $shelter_days$ and $covid_slackers$, have noticeably higher VIFs but are still below a rule-of-thumb level of concern (<5).

In [32]:
```
vif(v2model)
```

log(popdens): 1.74714248189299 **log(maxinsure):** 1.77446001748339 **elder:** 1.13292759971301 **close_biz_days:** 1.47047419055181 **shelter_days:** 3.4723826605274 **mask_days:** 1.47561138432999 **covid_slackers:** 3.56272745240808 **fedline:** 1.39893970153992 **northeast_corridor:** 1.75001492927819

3. CLM Assumptions Assessment for Model v1

Assumption 1

The model of the population is linear in parameters.

> We can see that the model is indeed linear in parameters. We have defined it as such. Even though the variables of interest have been transformed, the transformed quanities have the property of linearity as demonstrated by the scatterplot matrices.
>
> In terms of appropriateness, we see that the population is reasonably well represented by the linear modeling assumptions. That is to say, the errors are equally spread about the model as far as we can accurately measure them. It does not appear that the model would greatly improve its predictions of the population by being nonlinear.

Assumption 2

The observations used to create the model are a product of random sampling.

This assumption does NOT hold true for our data. We have a census, or at least an attempt at a census, not a sample. Because random sampling was not used, we can identify various biases (e.g. undercounting) and clustering effects (e.g. regional cross-state hotspots) in the data. The fact that the data doesn't meet this asusmption doesn't mean that the conclusions drawn from this study are worthless, but it does weaken the conclusions somewhat.

On the macro level, the individual rows compromise all 50 states plus DC. There are no states left to try to draw conclusions about. We might say that these represent a sample of all conceivable covid outcomes for states or that they represent a sample of all political entities over a certain size. However in both of these counts, they are a very biased sample. All recieved the same national guidance and existed in the same political climate. Individual groups of states even explicitly banded together to coordinate their covid responses in some cases.

On a more granular level, the infection and death counts are attempts at census that have inherent biases. There is strong reason to believe that both infections and deaths are being under-counted. In terms of infections, with few exceptions, only people who have strong symptoms are being tested. It is known that a significant portion of people who have the disease don't present with any symptoms. The only way to capture these asymptomatic people would be by random sampling. On top of that at the beginning of the crisis, there weren't enough tests even if you had severe symptoms in many cases. In terms of the death count, reporting standards vary from locality to locality. Some places would only count a covid death if the person had a positive test. Early in the crisis (perhaps still), they would not test people postumously because the tests are a finite resource. There's also evidence to suggest that early on, before the blood clotting symptom of covid was reported, a number of heart attacks that should have been associated with covid were not.

As noted above, all of these factors don't necessarily invalidate our conclusions, but they do weaken them. However, this is an important public event that needs to be understood so we are forced to use the data that we have.

UMN Report on Covid Deaths (https://www.cidrap.umn.edu/news-perspective/2020/07/about-30-covid-deaths-may-not-be-classified-such) Economist Report on Covid Excess Deaths (https://www.economist.com/graphic-detail/2020/07/15/tracking-covid-19-excess-deaths-across-countries)

Assumption 3

There is no perfect colinearity amongst the variables tested.

Looking at the v1 model, it can be seen that the unemployment benefit, the percentage of elderly residents, and the population density are not perfectly colinear. In fact, we can see from the correlation test in section 2.3.3 that they are very weakly related to each other in deed. All correlations are under 10%. The VIF shows that these low correlations don't translate into inflated variance for our regressors. This is good since even strong, but imperfect colinearity will make the models' estimates imprecise.

Assumption 4

The error u has an exepected value of zero.

We can see qualitatively from the residuals versus fitted plot output by `plot(v1model)` that the expected error is essentially zero over the range of the input data. There does not seem to be a consistent curve that is being neglected. Deviations from zero at the ends may be due to sparse data more than an actual trend. This means that zero conditional mean is satisfied in terms of our predictor variables. However, the chart cannot tell us if there is some larger endogeneity in our model that would cause the expected value of the error to be non-zero. We have to reason about that separately.

The three primary sources of endogeneity are (1) omitted variables, (2) reverse causality, (3) measurement error.

(1) Omitted variables are a real concern if they are stongly correlated with our variable of interest AND strongly correlated with the outcome variable. We have a detailed discussion of the size and possible effect of a variety of omitted variables in section 5 of this report. Unfortunately, several omitted variables that we came up with indidate that the effect of the unemployment payout may be smaller than we think. However, it should be noted that in the absence of those other omitted variables, the unemployment insurance payout might still be valuable as a proxy for those other more difficult to measure variables. We can still get predictive insights even if the causal link is weakened.

(2) We can rule out reverse causality. State unemployment insurance payouts were determined before covid struck the world.

(3) As discussed under assumption 2, there is reason to beleive that there is a measurement error where both deaths and cases are undercounted. However, if they are undercounted by a consitent percentage across the board (and do not vary with the chosen regressors, this should just manifest as an intercept change in our population model. The various importance of the variables would not be significantly affected.

Assumption 5

The error u has the same variance given any values of the explanatory variables.

In classical linear modeling, this assumption is needed to derive the standard errors of the estimator coefficients. We see from the scale versus location plot that this assumption largely holds true. The variance is a little larger towards the left and a little smaller towards the right which may indicate some amount of heteroskedasticity.

In this case we choose to use White standard error formulations that are robutst against heteroskedasticity. These will give acurate standard errors even when the residuals are not homoskedastic. This is conservative and probably warrented since it is not clear that the variables are completely homoskedastic.

Assumption 6

The population error is independent of the expanatory variables and is normally distributed.

This assumption is needed to determine the shape of the sampling distribution for our estimator coefficients. By looking at the Normal QQ plot of the model's residuals, we see that the model is almost normal but for some outliers in New York (which contains the most densely populated city in the country) and Wyoming (where the population is very spread out). When the shapiro wilke's test is run, we see that if the assumption of normality is true, we would only see a curve with deviations as extreme as those we see 4.96% of the time. This means that the errors are non-normal to a statistically significant degree but barely so.

Even if the errors are non-normal, a version of the CLM says that the sampling distribution for the estimator coefficients becomes normal as the number of observations increases. We have 51 observations which is more than the typical rule of thumb value of 30 where the CLM becomes applicable. Given the large nubmer of observations and the near-normality of the distribution. We can confidently apply the CLM in this case and the spirit of the requirement is satasfied.

See assumption 4 for a discussion of possible endogeneity.

4. Regression Table

For our model of choice (V1) we find that all of our variables of interest as well as the intercept are significant at the 0.05 level. We have similar results for V2, except that $elder$ has just been pushed out of the 10% significance window, likely due to collinearity introduced by our intentionally inclusive specification. (Probably for similar reasons, both of our other two key regressors are now only significant at the 10% level.)

Population density is the most significant variable in both V1 and V2, but perhaps most importantly we find that, once we control for obvious demographic effects, the key variable of our original research question $\log(maxinsure)$ is clearly significant.

While some of the government response covariates in V2 (close_biz_days and shelter_days) exhibit what seems to be reverse causality, their magnitude and significance are both low.

We somewhat unexpectedly found in V1 that the older a state's population is, the lower the infection rates as well. (The same finding is present in V2, but not statistically significant.) We think this is practically significant because of the correlation between old age and retirement in the population, and further illustrates the materiality of the need to work on infection rates.

Practically speaking, we see that for every 1% increase in the maximum unemployment insurance benefit, we find a 0.6% (for V0 and V1) or 0.8% (V2) decrease in covid-19 case rates per 100,000 people. We think this is practically significant given our stated research purpose, especially when compared to other policy-related variables.

In [33]:
```
stargazer(basemodel, v1model, v2model, type = "text", omit.stat = "f"
,
          se = list(se.basemodel, se.v1model, se.v2model),
          column.labels = c('Base', 'V1', 'V2'),
          star.cutoffs = c(0.1, 0.05, 0.01))
```

```
=================================================================
                                Dependent variable:
                    ---------------------------------------------
                                      log(infrate)
                          Base             V1              V2
                           (1)             (2)             (3)
-----------------------------------------------------------------
log(popdens)                            0.273***        0.235*
                                         (0.088)         (0.121)

log(maxinsure)          -0.583*         -0.586**        -0.835*
                        (0.348)          (0.281)         (0.433)

elder                                   -0.108**        -0.099
                                         (0.053)         (0.062)

close_biz_days                                           0.005
                                                        (0.008)

shelter_days                                             0.002
                                                        (0.009)

mask_days                                               -0.002
                                                        (0.004)

covid_slackers                                           0.538
                                                        (0.522)

fedline                                                  0.011
                                                        (0.050)

northeast_corridor                                       0.271
                                                        (0.435)

Constant                10.019***       10.578***       11.544***
                         (2.064)         (2.339)         (3.127)

-----------------------------------------------------------------
Observations               51              51              51
R2                        0.060           0.446           0.519
Adjusted R2               0.041           0.411           0.413
Residual Std. Error 0.736 (df = 49) 0.577 (df = 47) 0.576 (df = 41)
=================================================================
Note:                                  *p<0.1; **p<0.05; ***p<0.01
```

We're interested in a few other statistics as well. First, we look at the AIC for our improved and inclusive models to see which one has better fit.

In [34]:
```
print("AIC for v1 model:")
AIC(v1model)

print("AIC for v2 model:")
AIC(v2model)
```

```
[1] "AIC for v1 model:"

94.4888436265849

[1] "AIC for v2 model:"

99.2717413385374
```

It looks like our v1 model has a lower AIC and thus a better fit, but that result is partly by design. Recall that we deliberately put in quite a bit of regressors for our inclusive model, which the AIC penalizes via an increase in the $2k$ term.

When revisiting our v2 model, it's clear that further investigation is necessary to decide which regressors should stay in the model or not. To do so, we'll run a Wald test, which generalizes the usual F-test of overall significance but allows for a heteroskedasticity-robust covariance matrix.

What to use as our restricted model? We think a good case would be a v2 model without our "days of policy intervention" variables, i.e. 'close_biz_days', 'shelter_days', and 'mask_days'. These variables may well be jointly significant in capturing states' resolve in fighting coronavirus, but we aren't sure.

In [35]:
```
model_rest = lm(log(infrate) ~ log(popdens) + log(maxinsure) + elder
+ covid_slackers + fedline + northeast_corridor)

coeftest(model_rest, vcov = vcovHC)

waldtest(v2model, model_rest, vcov = vcovHC)
```

```
t test of coefficients:

                    Estimate Std. Error t value  Pr(>|t|)
(Intercept)       11.1319339  3.0184307  3.6880 0.0006181 ***
log(popdens)       0.2545094  0.1146926  2.2191 0.0316866 *
log(maxinsure)    -0.7497497  0.3987076 -1.8805 0.0666763 .
elder             -0.0923591  0.0590018 -1.5654 0.1246625
covid_slackers     0.4078180  0.2341608  1.7416 0.0885637 .
fedline            0.0091471  0.0475354  0.1924 0.8482926
northeast_corridor 0.2613924  0.3838411  0.6810 0.4994455
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

A anova: 2 × 4

	Res.Df	Df	F	Pr(>F)
	<dbl>	<dbl>	<dbl>	<dbl>
1	41	NA	NA	NA
2	44	-3	0.3032091	0.8228871

We fail to reject the null that the coefficients for our "days of polcy intervention" variables are equal to zero. This is a useful insight should we want to further develop model specifications, as it indicates the need to either eliminate these variables or further transform them. Perhaps a new variable taking the average of the three could be a parsimonious way to capture the effects of policy interventions.

5. Omitted Variables Discussion

We qualify our analysis by mentioning this is only a partial discussion: To really determine the true magnitude and direction of omitted variable bias on a multiple regression, we need to incorporate the covariance of every ommited variable with that of every regressor we've already included. To make this tractable, however, we assume that all other covariances are random and focus on our explanatory variable of interest only (maximum unemployment insurance payouts). Regardless of this assumption, we find these are the most relevant omitted variables and would include these even if we were to do an exhaustive discussion.

1. Cost of living

   Cost-of-living provisions are indexed in a number of benefits and negotiations such as workers' compensation, labor union contracts, or pensions; we expect unemployment insurance to be no different: Higher costs of living mean higher (nominal) unemployment insurance. In addition, higher costs of living mean a tighter squeeze if one is laid off, all else equal.

A higher cost of living increases the costs of not working, therefore it is positively correlated with case rates as workers find themselves needing to clock in.

We believe this bias to be (particularly) large and and towards zero.

While a built-in correlation also means that unemployment insurance could be a proxy for cost of living, this relationship would likely be weak as cost-of-living provisions are a complicated technical issue and decided at the state level.

1. % of jobs that can be done remotely Workers will be less likely to quit or be fired if their jobs can be or already are performed remotely (e.g. customer service workers, software developers). In addition, we observe that such jobs are on average higher-paid that manufactuting or agrarian jobs; higher-paying jobs are in turn correlated with higher unemployment insurance payouts.

We also believe this variable to negatively correlated with case rates - staying home puts is a straighforward risk-reducing activity.

As such, we estimate this bias to be moderate and away from zero.

We currently have no proxies for this in our dataset as is.

1. Time to receiving benefits (in crisis mode)

   Due to overwhelming demand, several states' unemployment insurance websites and/or systems crashed. As such, benefits were withheld for weeks or even months longer than usual; people counting on benefits to pay time-sensitive bills therefore likely had to head out in search of more work sooner than they should. We believe that the higher the unemployment insurance payout the more well-developed and robust a state's unemployment system, therefore the less time it would take to receive benefits.

We think that time to receiving benefits should be positively correlated with case rates, as people would seek work in lieu of unemployment insurance for immediate material needs.

These issues were eventually solved, so we believe the bias is small and biased away from zero.

We currently have no proxies for this in our dataset as is.

1. Average liquid assets

   One's wealth does not derive entirely from their wages - stocks, 401ks, savings, home equity, and other sources of wealth can be counted upon in a crisis. Some people will simply live off their rainy day funds instead of going straight back to work, and this variable would attempt to capture that. Higher average liquid assets probably correlate positively with unemployment insurance, due to the previously mentioned cost-of-living dynamic.

   Higher average liquid assets means one can wait out the pandemic, so we expect average liquid assets to be negatively correlated with case rates.

   We belive this bias to be (unfortunately) small due to the Americans' low savings rate and away from zero.

1. % Voters identifying as Republican

   Political affiliation would shape agent's behaviors in the current crisis. We could go for either party, but focus on Republicans' historical dislike of government intervention. We posit Republicans would be likelier to both stay open should there be no mandated closure *and* remain open in defiance of government mandates. In addition, we believe Republican sentiment correlates negatively with unemployment benefits, given the GOP's laissez-faire politics.

   Republican attitudes towards coronavirus have been anti-mask and generally skeptical of the virus' severity, so we believe this variable to be positively correlated with case rates.

   We believe this bias to be small and away from zero.

   We currently have no proxy for this in our dataset as is, although we could derive indicator variables using the 2016 or 2018 presidential/mid-term election results per state.

1. *Average* weekly unemployment insurance amount (dollars)

   This is a qualitatively different discussion, as we would rather have this as a replacement of our max insurance payout regressor. However, we still believe we can treat is an omittied variable.

   The weekly unemployment insurance *maximum* payout is not the most informative statistic possible. The *mean* would obviously be better - it captures the de facto unemployment insurance assistance a representative worker would receive. Average amounts are probably positively correlated with max amounts, as higher max amounts should be indicative of more generous unemployment systems.

   As motivated by our research question, we believe higher unemployment payouts (whether expressed through mean or max values) are negatively correlated with case rates.

   We believe the bias to be large and away from zero.

   Needless to say, our current unemployment insurance variable is a proxy, albeit an imperfect one.

5.1 Omitted Variables Summary

This sub-section contains a summary of our omitted variables discussion from above data is presented here in tabular form. Our primary resource for this sub-section was async section 10.14.

Let:

- $\boldsymbol{x_2}$ be the omitted variable under discussion
- $\boldsymbol{\delta_1}$ will have the same sign as the correlation of our omitted variable with our regressor of interest, ie. ln(maxinsure).
- $\boldsymbol{\beta_2}$ be the hypothesized relationship between the ommitted variable (x_2) and our y variable, i.e. the confirmed coronavirus cases (y). A '+' indicates that the relationship is positive (a higher x_2 will lead to a higher y). A '-' indicates that the relationship is negative (a higher x_2 will lead to a lower y).
- **Net Effect** be the net directional effect of the omitted variable bias.

Where $\alpha_1 = \beta_1 + \beta_2\ \delta_1$ is the biased coefficient and $\beta_2\ \delta_1$ is the bias itself.

When the net effect is negative, the bias is away from zero; when the net effect is positive the bias is towards zero.

Omitted Variable (x_2)	δ_1	β_2	**Net Effect**
Cost of living	+	+	+
% Jobs that can be remote	+	-	-
Time to (crisis-mode) benefit receipt	-	+	-
Avg. liquid assets	+	-	-
% Republican voters	-	+	-
Avg unemployment payout	+	-	-

6. Conclusion

Our independent variable of choice has shown statistical significace in explaining COVID infection rates across the US, highlighting the importance of the *economic* channel when discussing how to best address infection rates. This statistical significance holds at the 5% level at our improved specifications and the 10% level in our baseline and inclusive model, underscoring its robustness. Its large coefficient versus other covariates also illustrates its practical significance: By way of example, our model estimates that (ceteris paribus) a 1% increase in the max unemployment insurance payout for California leads to about 1,500 less COVID cases in the state.

While demographics certainly matter, only so much can be done about living in a crowded city or having an elderly population. In addition, government responses are effective insofar as they are enforced and (most crucially) *complied with*. in a world of lockdowns and prohibitions, a good social safety net is a key way to gain popular goodwill while ensuring government mandates aren't undermined by people's need to work.

Unemployment insurance is only one dimension of economic aid, and a temporary one at that - other, more innovative subsidies and incentives will be crucial going forward if we are serious about stopping coronavirus, be it this wave or the next ones.

In [36]:
```
# References - visuals
# Intro visual - https://www.npr.org/sections/coronavirus-live-updates/2020/03/26/821933358/advice-on-filing-for-unemployment-benefits-document-everything-and-be-persistent
# "Persons tested positive by county" visual -- https://covid19tracker.health.ny.gov/views/NYS-COVID19-Tracker/NYSDOHCOVID-19Tracker-Map?%3Aembed=yes&%3Atoolbar=no&%3Atabs=n
# NYC visuals for cases by age, poverty: https://www1.nyc.gov/site/doh/covid/covid-19-data.page (data taken Aug. 5)
# Acela Corridor - https://en.wikipedia.org/wiki/Acela
# Rule of thumb value/discussion for VIF - there is no *one* rule, but we found this discussion most appropriate - https://www.researchgate.net/post/Multicollinearity_issues_is_a_value_less_than_10_acceptable_for_VIF
# Dataset: Maki-Nayeri, Majid et al. “COVID-19 W203 Lab 3 Database.” UC Berkeley , 8 July 2020.
```