Lab 2: Epidemic Growth

First, read in data from the New York Times.

fulldata<-read.csv('https://raw.githubusercontent.com/nytimes/covid-19-data/master/us-counties.csv')  
head(fulldata)

## date county state fips cases deaths  
## 1 2020-01-21 Snohomish Washington 53061 1 0  
## 2 2020-01-22 Snohomish Washington 53061 1 0  
## 3 2020-01-23 Snohomish Washington 53061 1 0  
## 4 2020-01-24 Cook Illinois 17031 1 0  
## 5 2020-01-24 Snohomish Washington 53061 1 0  
## 6 2020-01-25 Orange California 6059 1 0

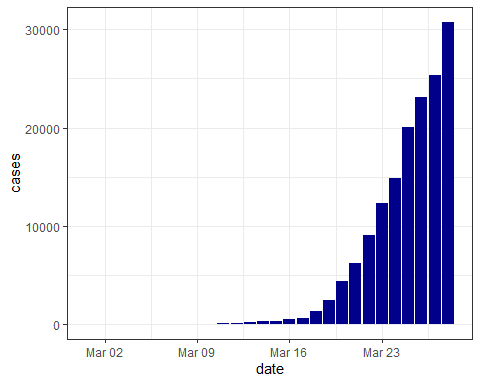
Select the data from New York City using **tidyverse**.

library(tidyverse)  
  
#this is how to manipulate data using tidyverse 'pipes'. 'filter' selects observations. 'mutate' creates new variables. we have to define the date variable as being a date with format Year-month-day.  
data <- fulldata %>%  
 filter(county=="New York City") %>%  
 mutate(date= as.Date(date,"%Y-%m-%d"))

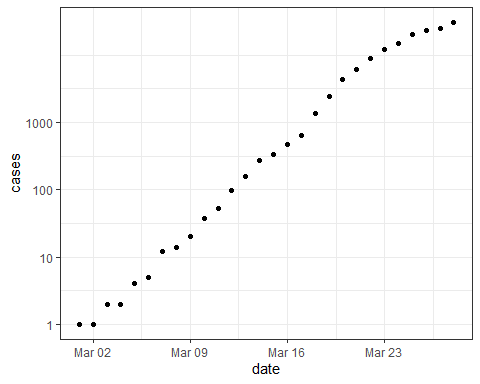
## Estimating the Growth Rate

Start by visualizing the number of cases in the first 4 weeks, both on the regular and log scales.

#first, create a dataset containing just the observations for the first 4 weeks  
data.early <- data %>%  
 filter(date >= as.Date("2020-03-01","%Y-%m-%d")) %>%  
 filter(date <= as.Date("2020-03-28","%Y-%m-%d"))   
  
#create a bar plot  
data.early %>%  
 ggplot(aes(x=date,y=cases)) + geom\_bar(stat='identity',fill='darkblue') + theme\_bw()

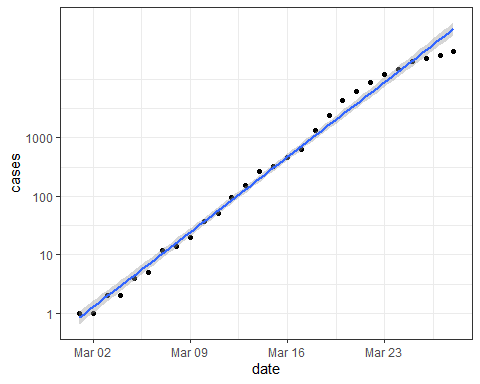


#plot the cases on the log scale  
data.early %>%  
 ggplot(aes(x=date,y=cases)) + geom\_point() + theme\_bw() + scale\_y\_continuous(trans='log', breaks=c(1,10,100,1000))



It looks mostly exponential (linear on the log scale), with some slowing of growth towards the end of March. Let’s plot a regression line and see.

#here we plot again on log scale but use 'stat\_smooth' with a linear model ('lm')  
data.early %>%  
 ggplot(aes(x=date,y=cases)) + geom\_point() + theme\_bw() +  
 stat\_smooth(method = "lm", formula = y ~ x) + scale\_y\_continuous(trans='log', breaks=c(1,10,100,1000))



There appears to be a good fit, except towards the end of March as predicted. Inspect the regression more closely to extract the estimated growth rate.

#run a linear regression with outcome of log(cases) and time (date) as the predictor  
model1 <- lm(log(cases) ~ date, data=data.early)  
summary(model1)

##   
## Call:  
## lm(formula = log(cases) ~ date, data = data.early)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.8751 -0.1482 0.0139 0.1892 0.5572   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -7.731e+03 1.433e+02 -53.95 <2e-16 \*\*\*  
## date 4.219e-01 7.816e-03 53.98 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.3341 on 26 degrees of freedom  
## Multiple R-squared: 0.9912, Adjusted R-squared: 0.9908   
## F-statistic: 2914 on 1 and 26 DF, p-value: < 2.2e-16

Here, the growth rate (slope) is the regression coefficient for ‘date’, which is 4.219e-01 (i.e. 0.42 or 42% per day). The coefficient is statistically significant and R-squared is high.

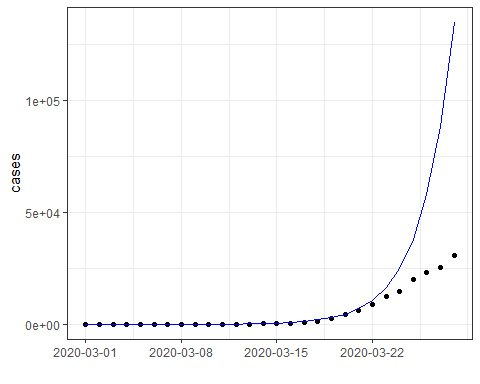
Note: if you had changed the start date in this part of the analysis to March 7, you would have gotten a slightly lower growth rate (around 41%)

Let’s examine the fit on the regular scale.

#the growth rate is the coefficient for date  
growthrate= coef(model1)[2]  
growthrate

## date   
## 0.4219154

#start with the initial number of cases in the series  
initialcases= data.early$cases[1]  
  
ggplot() +   
 geom\_point(data=data.early,aes(x=1:28,y=cases)) +  
 geom\_line(aes(x=1:28,y=initialcases\*exp((1:28)\*growthrate)),color='blue') + theme\_bw() +   
 scale\_x\_continuous(breaks=seq(1,28,7),labels=seq(as.Date('2020-03-01'), as.Date('2020-03-28'), by = "7 days")) + xlab("")



The fit from about the last week of March on is quite poor.

## Estimating the Reproductive Number ()

Now that we have the growth rate, we can estimate , assuming we know the latent and infectious periods. Let’s assume that the latent period (L) is 3 days and the infectious period (D) is 4 days. Now, we can use the equation:

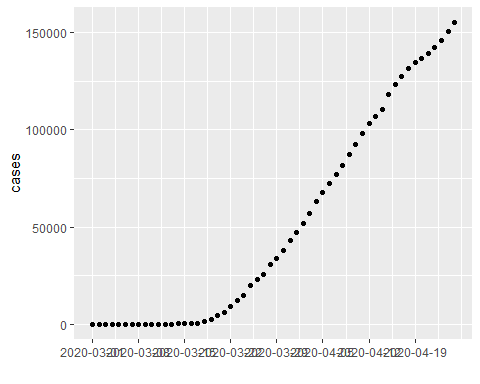
L=3  
D=4  
r= as.numeric(coef(model1)[2]) #0.42  
R0= (1+L\*r)\*(1+D\*r)  
print(paste0("R0: ",round(R0,2)))

## [1] "R0: 6.09"

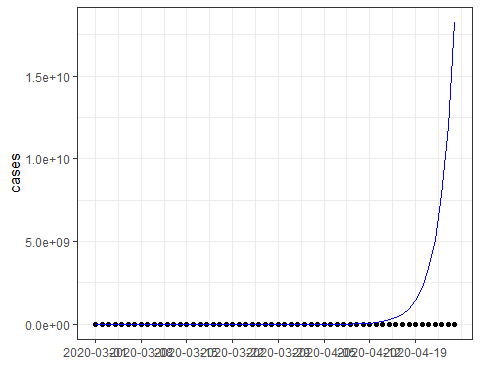
Note: if you had changed the start date in this part of the analysis to March 7, you would have gotten a slightly lower R0 (closer to 5.9)

Did NYC’s epidemic continue to have this same reproductive number, and therefore growth rate, in April? If so, how many cases would it have had?

#plot just the actual cases to start  
data %>%   
 filter(date >= as.Date("2020-03-01","%Y-%m-%d")) %>%  
 filter(date <= as.Date("2020-04-25","%Y-%m-%d")) %>%  
 ggplot() + geom\_point(aes(x=1:56,y=cases)) +  
 scale\_x\_continuous(breaks=seq(1,56,7),labels=seq(as.Date('2020-03-01'), as.Date('2020-04-25'), by = "7 days")) + xlab("")

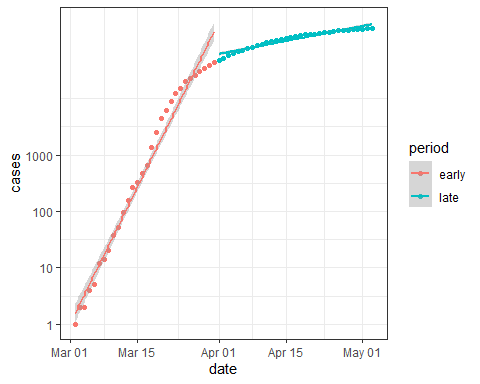


#here we plot the cases along with the exponential growth from the first four weeks, projected forward  
data %>%   
 filter(date >= as.Date("2020-03-01","%Y-%m-%d")) %>%  
 filter(date <= as.Date("2020-04-25","%Y-%m-%d")) %>%  
 ggplot() + geom\_point(aes(x=1:56,y=cases)) +  
 geom\_line(aes(x=1:56,y=initialcases\*exp((1:56)\*growthrate)),color='blue') + theme\_bw() +   
 scale\_x\_continuous(breaks=seq(1,56,7),labels=seq(as.Date('2020-03-01'), as.Date('2020-04-25'), by = "7 days")) + xlab("")

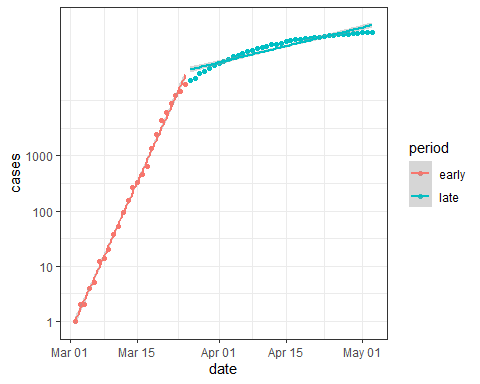


The fit is quite poor, and from the last week of March growth looks more linear than exponential. Let’s look at cases on the log scale, and estimate the growth rate for March versus April. We notice that it may be better to start the “late” period in the last week of March rather than April, which could fix the issues we had earlier.

#we define and early and late period, and then plot them in pieces, with regression line in pieces.  
#using original early and late periods  
data %>%  
 filter(date > as.Date("2020-03-01","%Y-%m-%d")) %>%  
 mutate(period = case\_when(date <= as.Date("2020-03-31","%Y-%m-%d") ~ 'early',  
 date > as.Date("2020-03-31","%Y-%m-%d") ~ 'late')) %>%  
 ggplot(aes(x=date,y=cases,color=period)) + geom\_point() + theme\_bw() + scale\_y\_continuous(trans='log',breaks=c(1,10,100,1000)) +   
 stat\_smooth(method = "lm", formula = y ~x)



#using new early and late periods  
data %>%  
 filter(date > as.Date("2020-03-01","%Y-%m-%d")) %>%  
 mutate(period = case\_when(date <= as.Date("2020-03-25","%Y-%m-%d") ~ 'early',  
 date > as.Date("2020-03-25","%Y-%m-%d") ~ 'late')) %>%  
 ggplot(aes(x=date,y=cases,color=period)) + geom\_point() + theme\_bw() + scale\_y\_continuous(trans='log',breaks=c(1,10,100,1000)) +   
 stat\_smooth(method = "lm", formula = y ~x)



The second graph (with the “later” period of slow growth starting earlier) looks better, although growth may still be too slow to be exponential (less than a linear increase on the log scale).Let’s calculate the for April (to date) and compare it to what we estimated for the the first month and for the revised early period. First we’ll fit a regression of log cases against time to estimate the growth rate. Then we’ll again use our equation for as above.

#original early and late periods  
data.late <- data %>%  
 filter(date >= as.Date("2020-04-01","%Y-%m-%d")) %>%  
 filter(date <= as.Date("2020-04-25","%Y-%m-%d"))   
  
model2 <- lm(log(cases) ~ date, data=data.late)  
L=3  
D=4  
r= as.numeric(coef(model2)[2])  
print(paste0("r: ",round(r,2)))

## [1] "r: 0.05"

R0\_late= (1+L\*r)\*(1+D\*r)  
print(paste0("Original Early R0: ", round(R0, 2)))

## [1] "Original Early R0: 6.09"

print(paste0("Original Late R0: ",round(R0\_late,2)))

## [1] "Original Late R0: 1.36"

#new early and late periods  
data.early <- data %>%  
 filter(date >= as.Date("2020-03-01","%Y-%m-%d")) %>%  
 filter(date <= as.Date("2020-03-25","%Y-%m-%d"))   
  
model2 <- lm(log(cases) ~ date, data=data.early)  
r= as.numeric(coef(model2)[2])  
print(paste0("r: ",round(r,2)))

## [1] "r: 0.44"

R0\_early= (1+L\*r)\*(1+D\*r)  
  
data.late <- data %>%  
 filter(date >= as.Date("2020-03-26","%Y-%m-%d")) %>%  
 filter(date <= as.Date("2020-04-25","%Y-%m-%d"))   
  
model2 <- lm(log(cases) ~ date, data=data.late)  
r= as.numeric(coef(model2)[2])  
print(paste0("r: ",round(r,2)))

## [1] "r: 0.06"

R0\_late= (1+L\*r)\*(1+D\*r)  
print(paste0("New Early R0: ", round(R0\_early, 2)))

## [1] "New Early R0: 6.44"

print(paste0("New Late R0: ",round(R0\_late,2)))

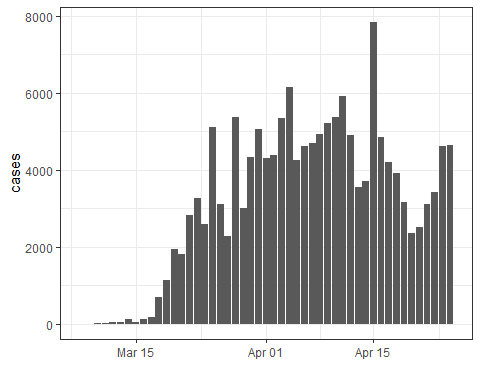
## [1] "New Late R0: 1.46"

The reproductive number for April was about 1/4 of that in March. And even that appears to be an overestimate.

## Estimating a Continuous Reproductive Number ()

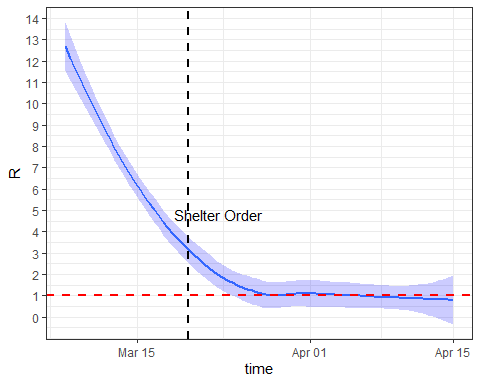
What we would really like is a continuous measure of the reproductive number (). This is what the Wallinga-Teunis method delivers. Let’s apply that to the NYC data. First let’s plot new cases by day.

#define the first date of the series  
first.date<- as.Date("2020-03-07","%Y-%m-%d")  
  
library(R0)  
  
#create a dataset from the first date through April 25  
data.march.april <- data %>%  
 filter(date >= first.date) %>%  
 filter(date <= as.Date("2020-04-25","%Y-%m-%d"))   
  
#we need to transform cumulative cases into incidence  
case.inc<- check.incid(diff(data.march.april$cases), date.first.obs=first.date+1, time.step=1)  
  
ggplot() + geom\_bar(aes(x=case.inc$t,y=case.inc$incid),stat='identity') + theme\_bw() + ylab('cases') +  
 xlab('')



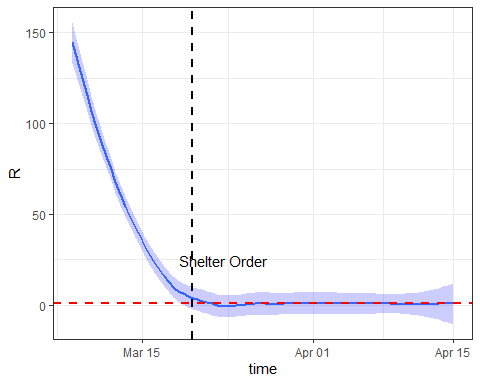
Now let’s plot the .

mGT<-generation.time("gamma", c(5.8, 3)) #this is the generation interval, defined as a gamma distribution with a mean of 5.8 days  
R0.res<-est.R0.TD(epid=case.inc$incid,GT=mGT)  
  
result <- data.frame(R= R0.res$R, lowCI= R0.res$conf.int[,1], highCI= R0.res$conf.int[,2],  
 time= case.inc$t[1:length(R0.res$R)])  
  
ggplot(result,aes(x=time,y=R)) + geom\_smooth(aes(ymin=lowCI,ymax=highCI),fill='blue',alpha=0.2) + theme\_bw() + geom\_hline(yintercept=1, linetype="dashed", color = "red", size=1) +   
 geom\_vline(xintercept=as.Date("2020-03-20","%Y-%m-%d"),linetype='dashed',color='black',size=1) +  
 annotate("text",x=as.Date("2020-03-23","%Y-%m-%d"),y=4.8,label='Shelter Order') +  
 scale\_y\_continuous(breaks=c(0:100))

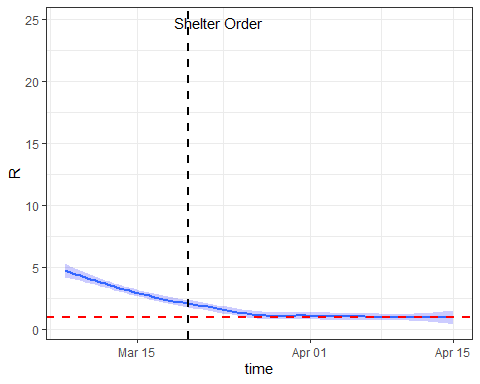


Now we change the generation interval, first doubling it from 5.8 days to 11.6 days

mGT<-generation.time("gamma", c(11.6, 3)) #this is the generation interval, defined as a gamma distribution with a mean of 5.8 days  
R0.res<-est.R0.TD(epid=case.inc$incid,GT=mGT)  
  
result2 <- data.frame(R= R0.res$R, lowCI= R0.res$conf.int[,1], highCI= R0.res$conf.int[,2],  
 time= case.inc$t[1:length(R0.res$R)])  
  
ggplot(result2,aes(x=time,y=R)) + geom\_smooth(aes(ymin=lowCI,ymax=highCI),fill='blue',alpha=0.2) + theme\_bw() + geom\_hline(yintercept=1, linetype="dashed", color = "red", size=1) +   
 geom\_vline(xintercept=as.Date("2020-03-20","%Y-%m-%d"),linetype='dashed',color='black',size=1) +  
 annotate("text",x=as.Date("2020-03-23","%Y-%m-%d"),y=24.8,label='Shelter Order') +  
 scale\_y\_continuous()



mGT<-generation.time("gamma", c(2.9, 3)) #this is the generation interval, defined as a gamma distribution with a mean of 5.8 days  
R0.res<-est.R0.TD(epid=case.inc$incid,GT=mGT)  
  
result3 <- data.frame(R= R0.res$R, lowCI= R0.res$conf.int[,1], highCI= R0.res$conf.int[,2],  
 time= case.inc$t[1:length(R0.res$R)])  
  
ggplot(result3,aes(x=time,y=R)) + geom\_smooth(aes(ymin=lowCI,ymax=highCI),fill='blue',alpha=0.2) + theme\_bw() + geom\_hline(yintercept=1, linetype="dashed", color = "red", size=1) +   
 geom\_vline(xintercept=as.Date("2020-03-20","%Y-%m-%d"),linetype='dashed',color='black',size=1) +  
 annotate("text",x=as.Date("2020-03-23","%Y-%m-%d"),y=24.8,label='Shelter Order') +  
 scale\_y\_continuous()



## Homework

For homework:

* Redo the above analyses, but instead for New York City (county “New York City”). What were the early growth rate and in New York City? *Answer: The early growth rate was 42% (much higher than the 16% we observed in Santa Clara County in class) and the was 6.09. However, there is a poor fit towards the end of March, where the exponential growth model predicts far more cases than were actually observed.*
* The analysis will not run, so try changing the start date (variable first.date) to March 7. Note that the shelter order for New York City began on March 20 rather than March 17
* What would happen if we assumed a longer or shorter serial (generation) interval? Try doubling the generation time (currently 5.8 days) or making it half as long. What does this do to ? Describe the change. *Answer: With a slower moving epidemic (longer serial/generation interval), to observe what we see at the beginning of March in NYC we’d have to have a much higher . With a faster moving epidemic (shorter serial interval), would have to be much lower to observe this level of growth. By April, once growth has slowed,the ’s are much more similar, with overlapping confidence intervals. This indicates that growth is less sensitive to the generation interval for a disease with a lower , which is confirmed via the equations from class as well.*