Analysis of Publicly Available Covid-19 Data

KENNESAW STATE UNIVERSITY ANALYTICS SHOWCASE

MARCH 3, 2021

CONNOR ARMSTRONG AND DR. AUSTIN BROWN

KENNESAW STATE UNIVERSITY

Topics

1. Visualizing global trends of number of reported cases and vaccinations over time
2. Investigation into the relationship between reported cases and vaccines

Source / Data Structure

Our World in Data https://ourworldindata.org/coronavirus-source-data

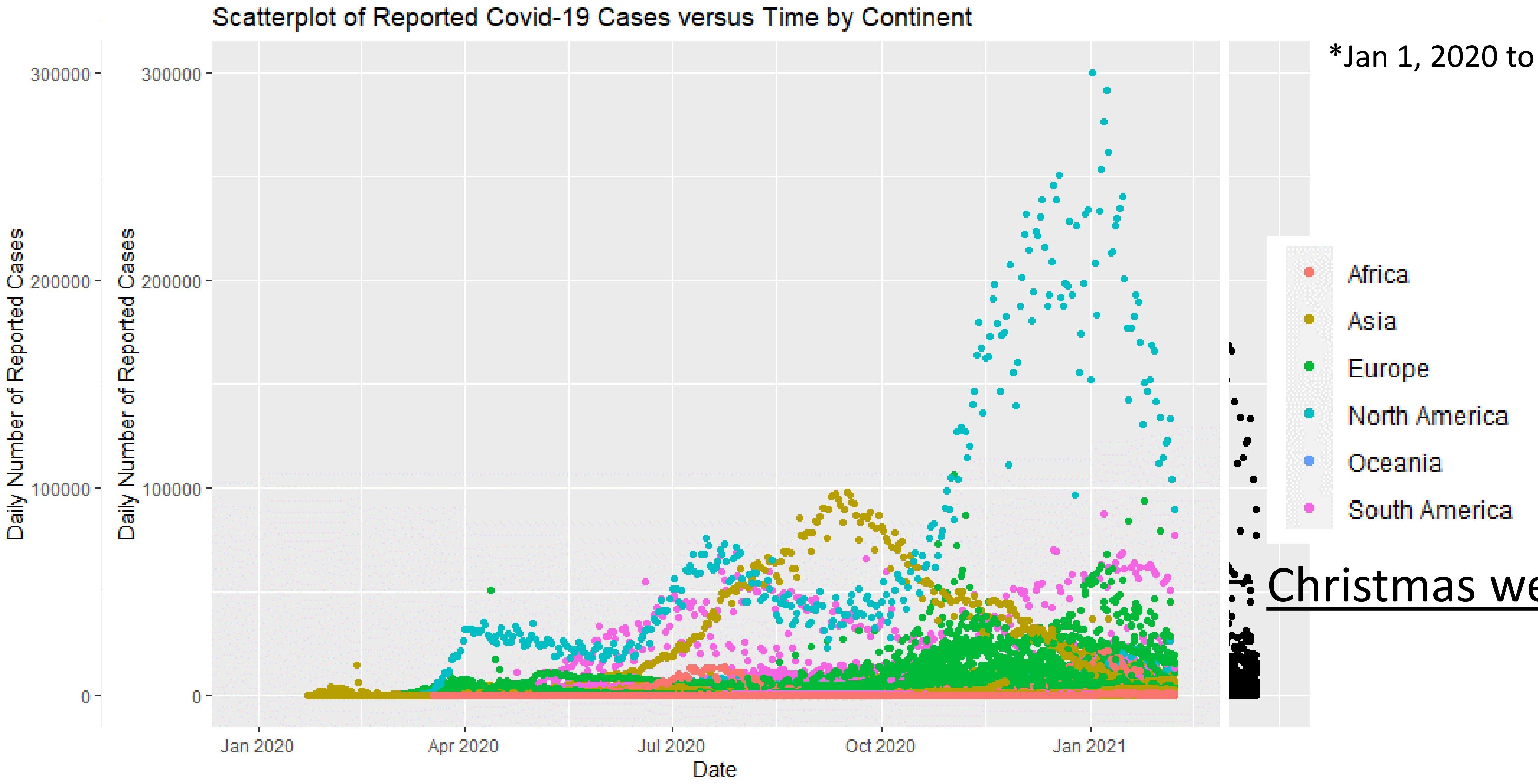
From the ECDC (European Center for Disease Prevention and Control)

Daily reportings of new cases, deaths, vaccines, and tests by region

59 variables

KENNESAW STATE UNIVERSITY

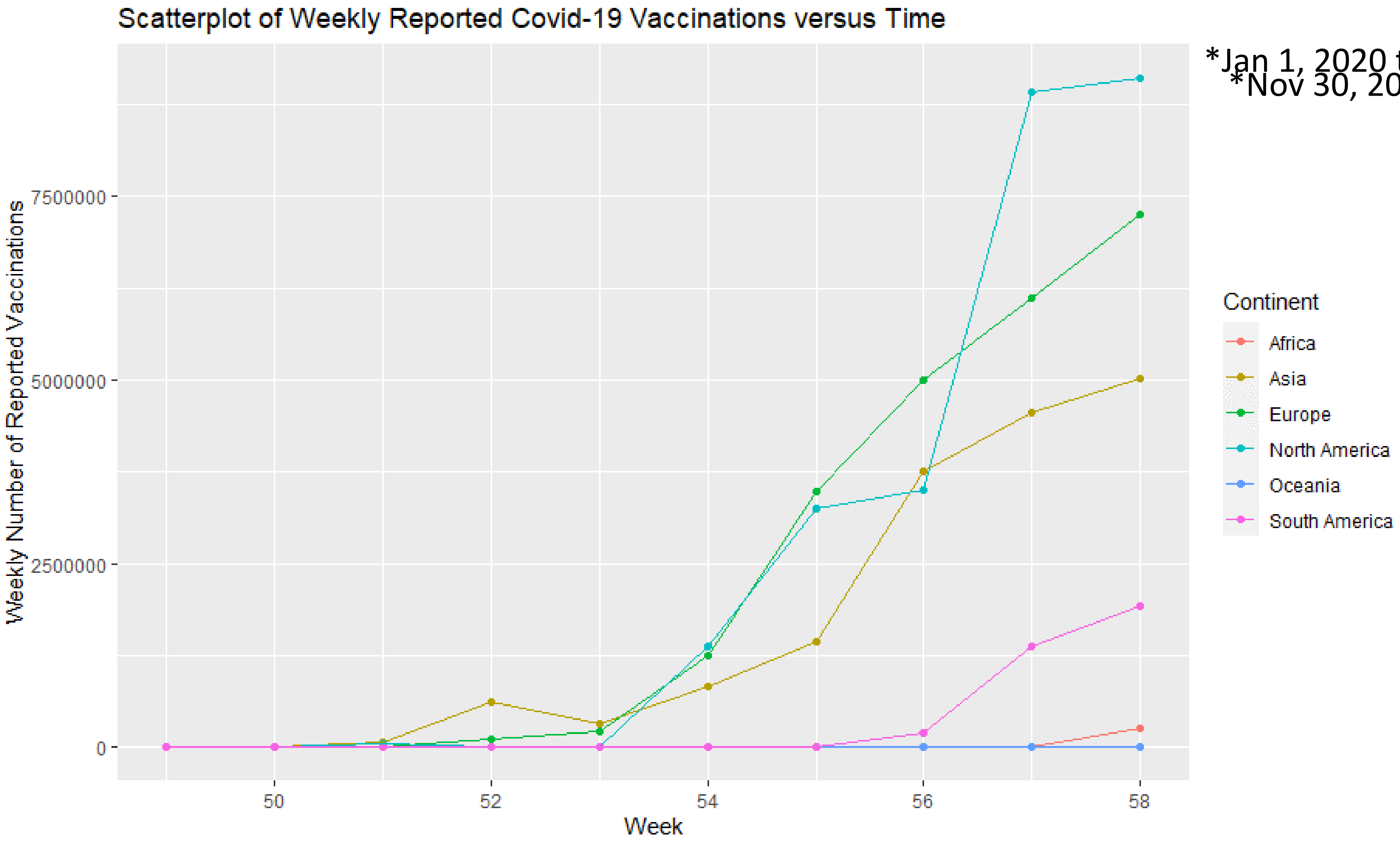
Reported Covid-19 Cases Over Time



*Jan 1, 2020 to Feb 8, 2021

Christmas week!

Reported Covid-19 Vaccinations Over Time

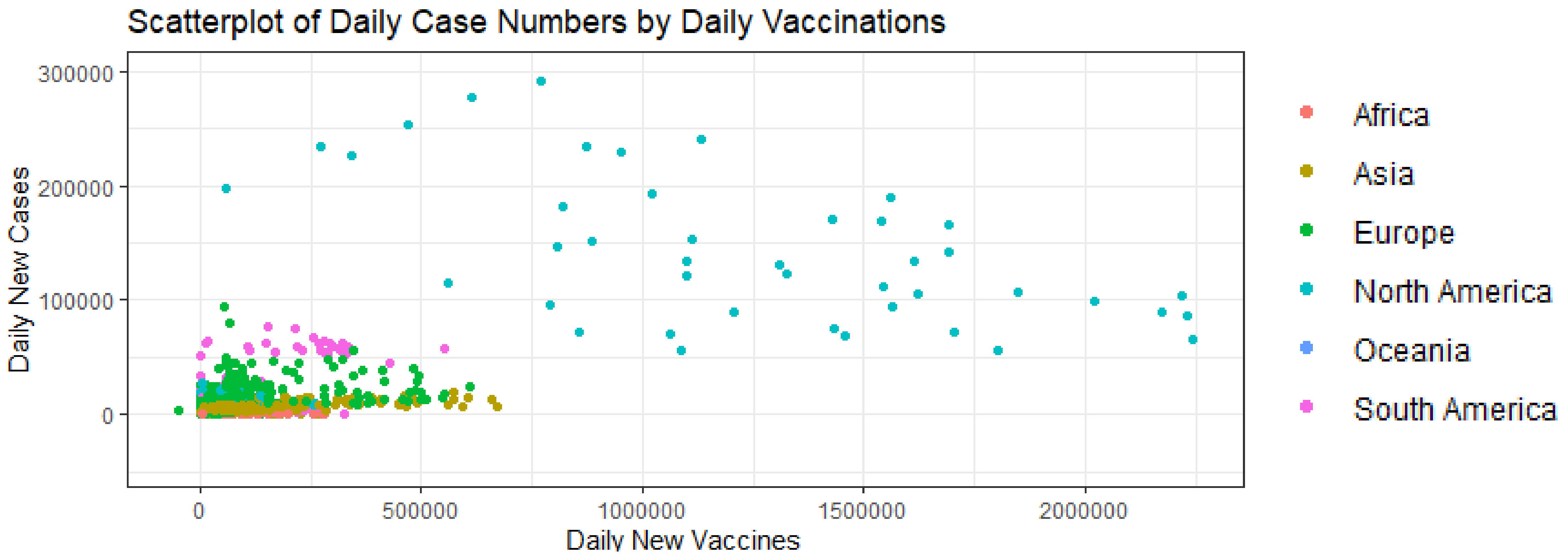


*Jan 1, 2020 to Feb 8, 2021
*Nov 30, 2020 to Feb 8, 2021

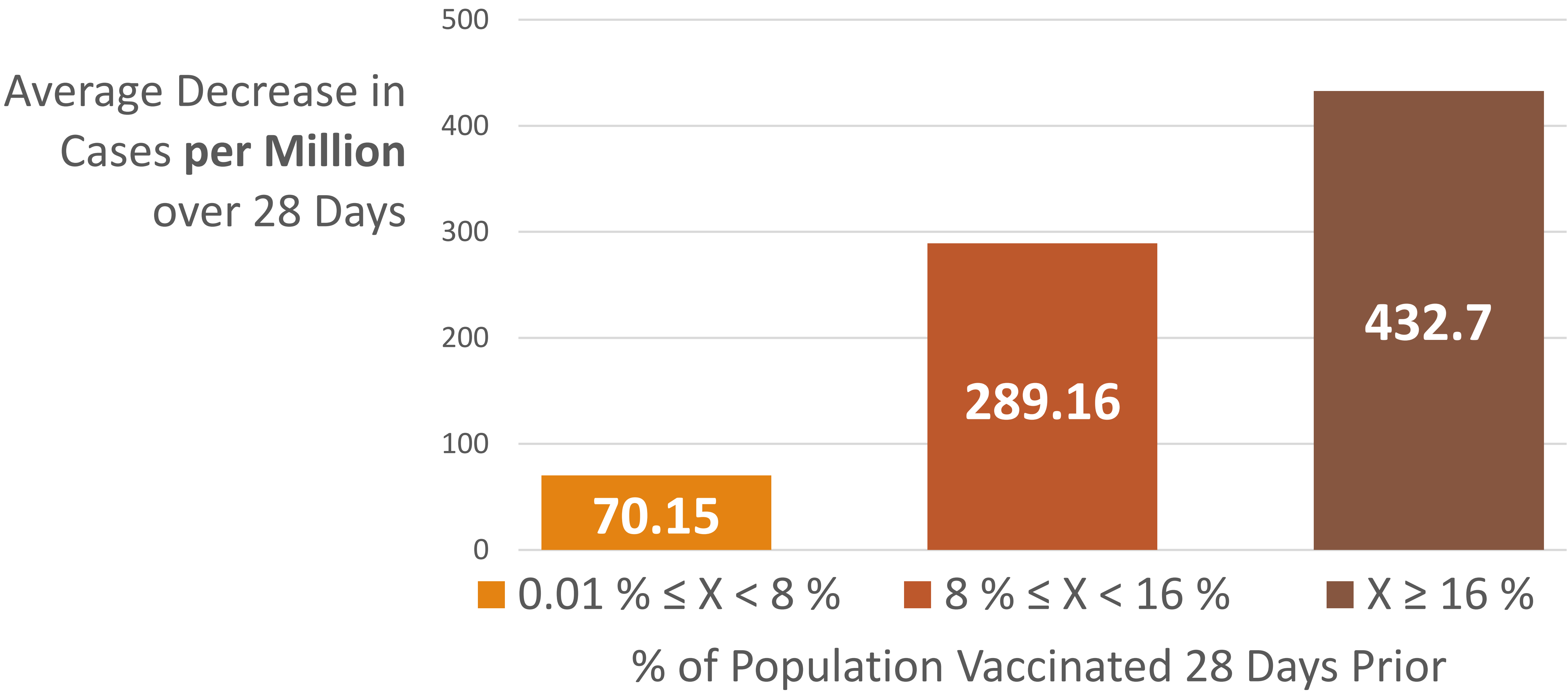
How do vaccines affect case numbers?

The following issues make it difficult to answer this question:

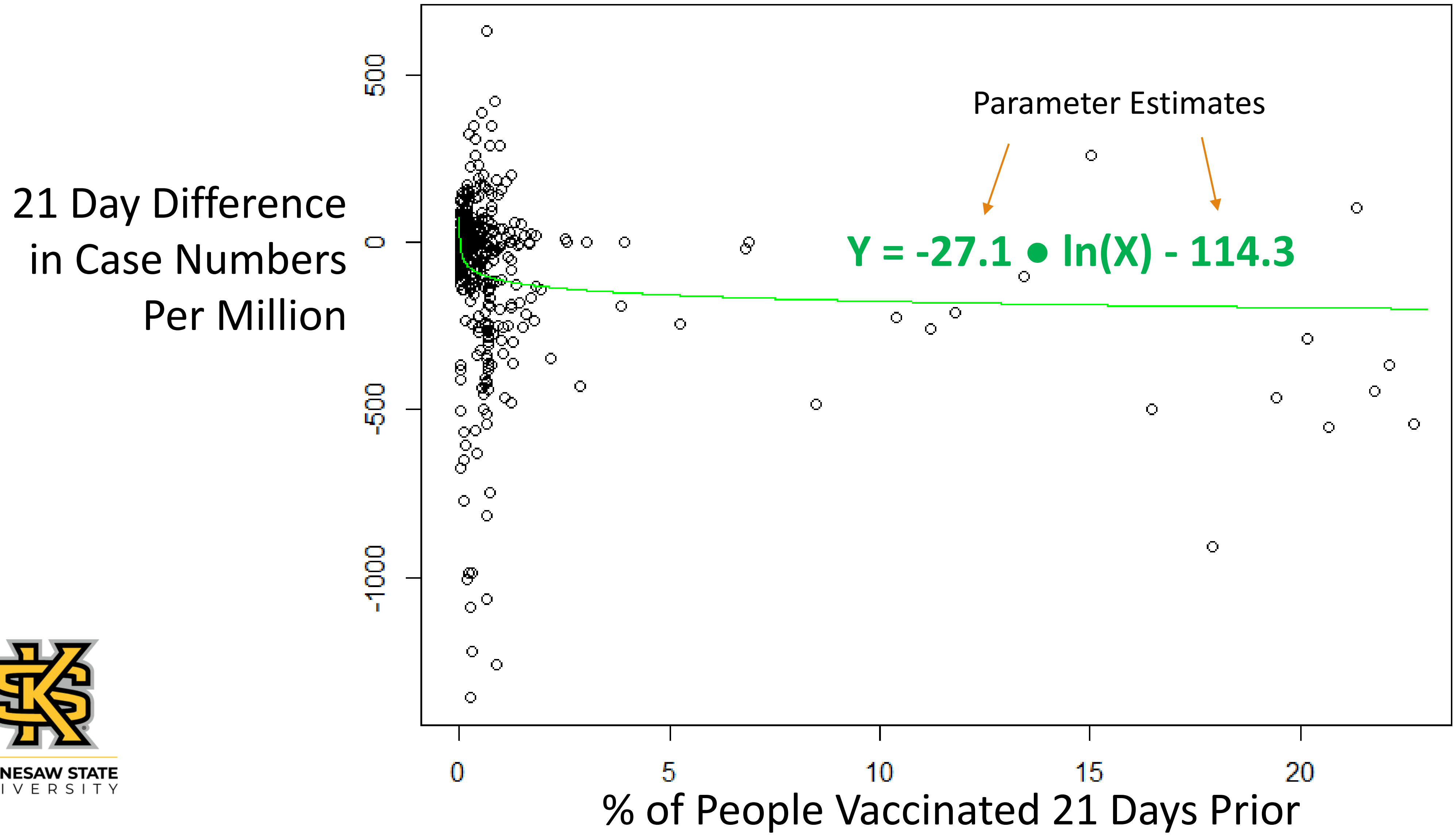
PROBLEM		SOLUTION
No clear relationship in cases and vaccines	→	Use differences from past days
Scale of case and vaccine numbers vary significantly	→	Use cases per million and % vaccinated



More People Vaccinated Results in Larger Decreases in Case Numbers

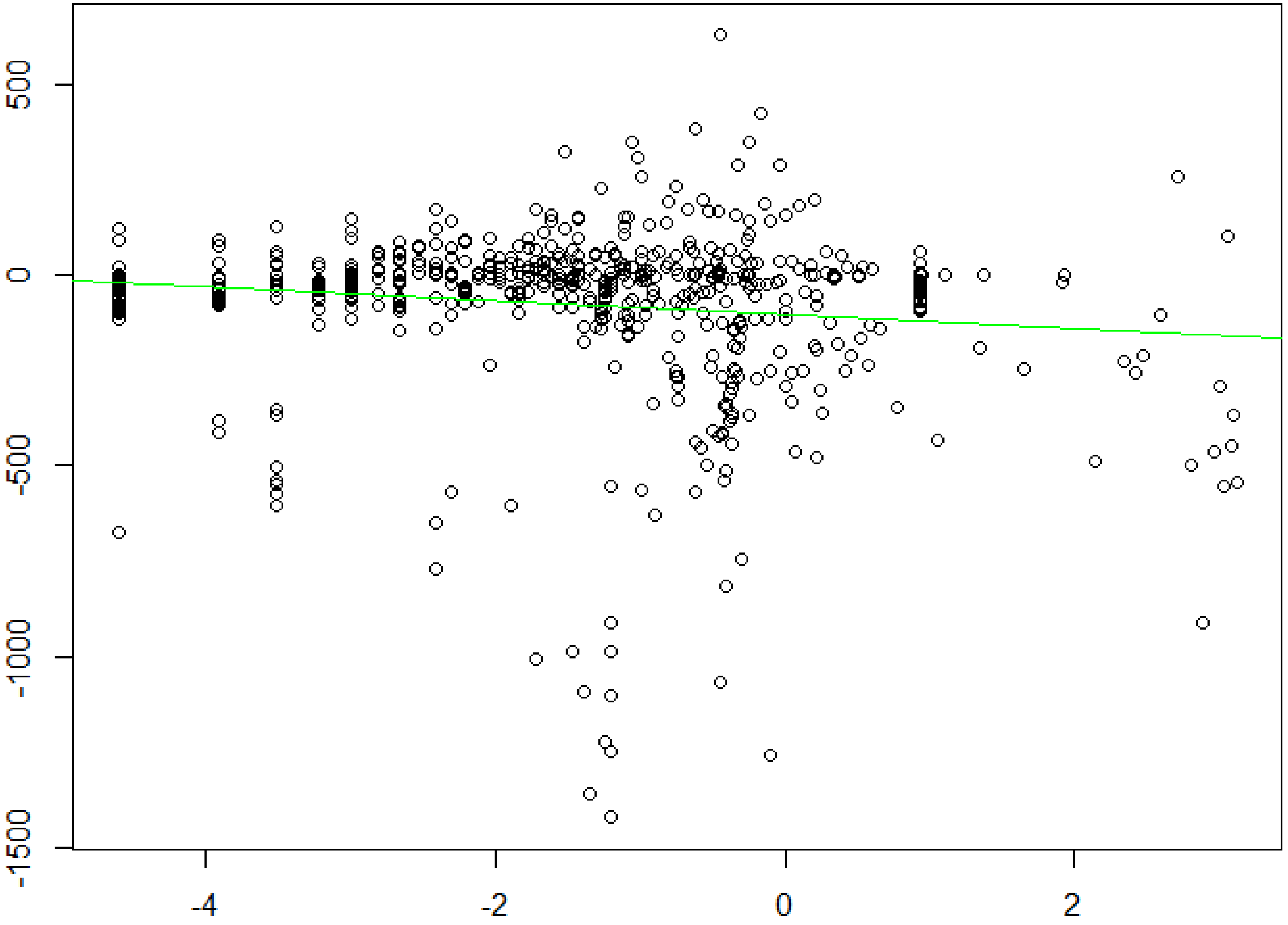


Best Fit Model for 21 Day Differences



KENNESAW STATE UNIVERSITY

Best Fit Model for 21 Day Differences, Log Scale

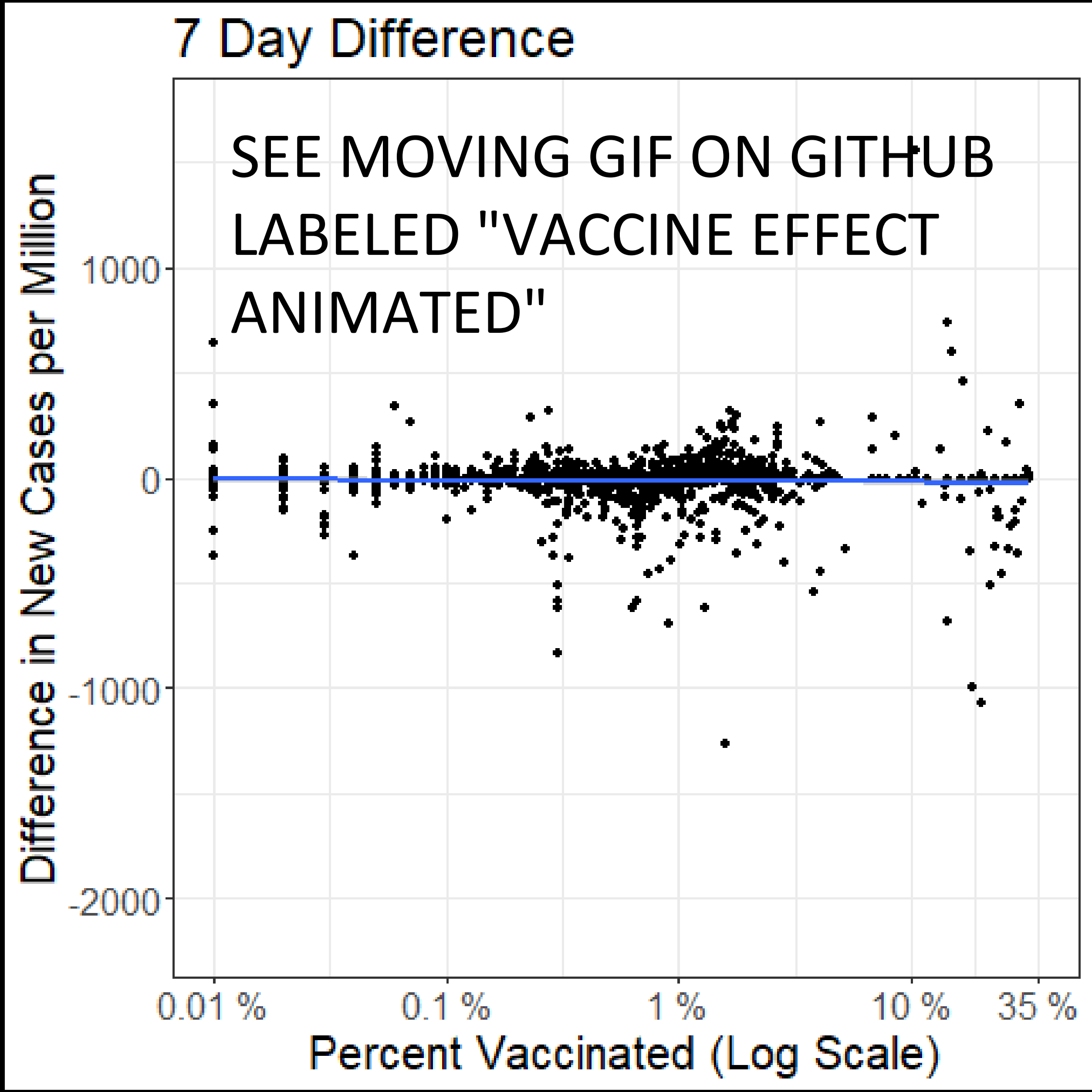


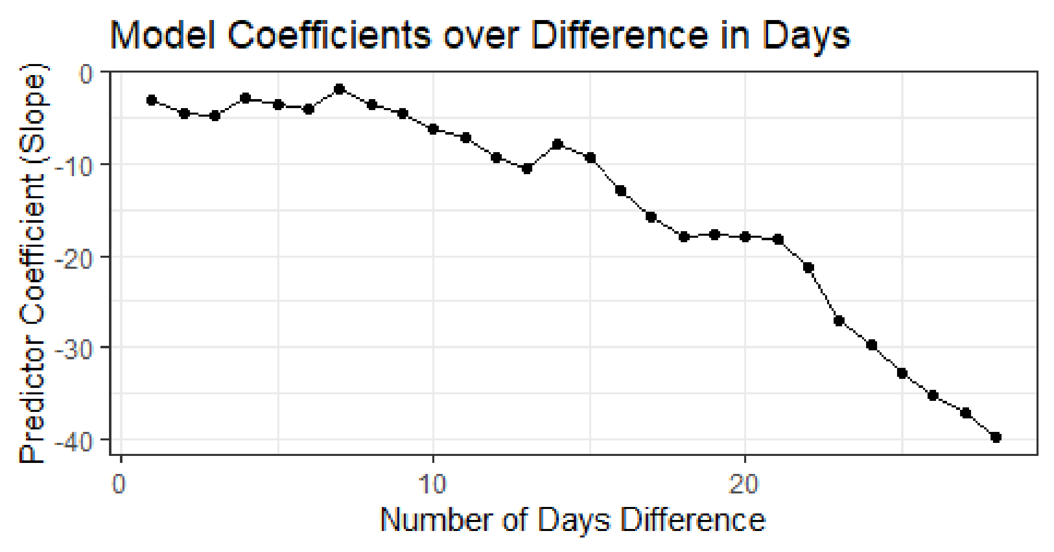
21 Day Difference in Case Numbers Per Million

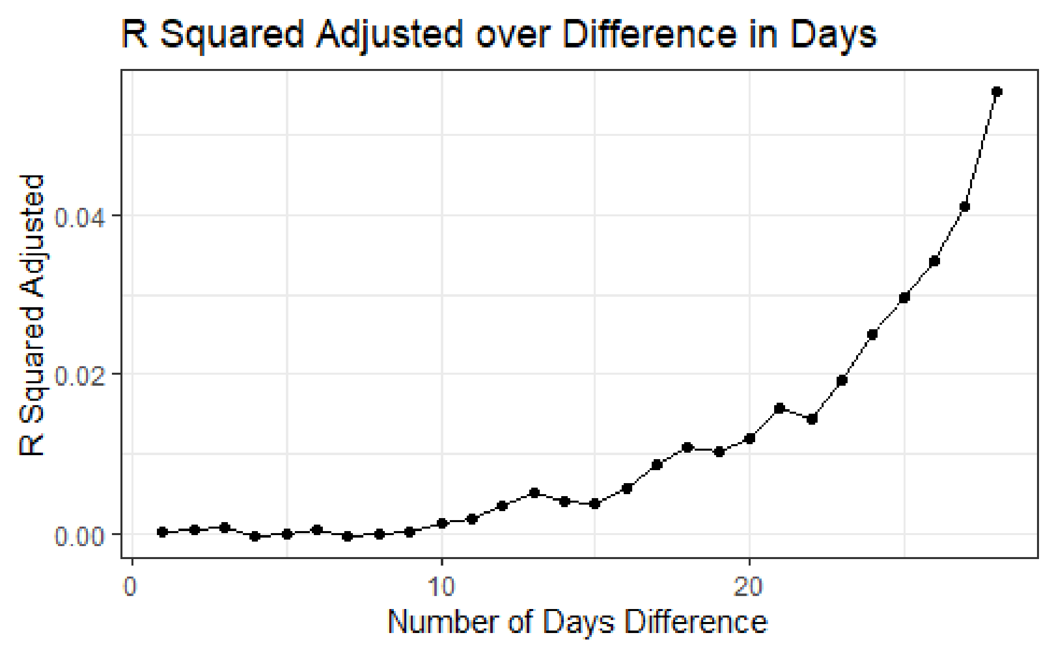
Natural Log of (% of People Vaccinated 21 Days Prior)

KENNESAW STATE UNIVERSITY

How Does Percent Vaccinated Affect the Difference in Case Numbers Over Time?







Summary

- Sometimes relationships between phenomena are relative.
- Countries which report higher percentages of people vaccinated appear to be more likely to see larger decreases in daily case numbers in the future.
- Increasing the number of days after vaccination increases the magnitude of its effect on case numbers.

Questions?

Appendix

Interpretation of model coefficients

<u>Predictor Estimate</u> – units are **(Difference in Case Numbers)** / **ln(% Vaccinated)**

For every 1 unit increase in **ln(% vaccinated)** there is an X **Change in Case Numbers** predicted

<u>Intercept Estimate</u> – unit is **difference in case numbers**

The predicted difference in case numbers when **ln(% vaccinated) = 0 = ln(1) <-> e^0 = 1**

In other words, 1 % is 1, not 0.01

Vaccinations are a recent phenomena

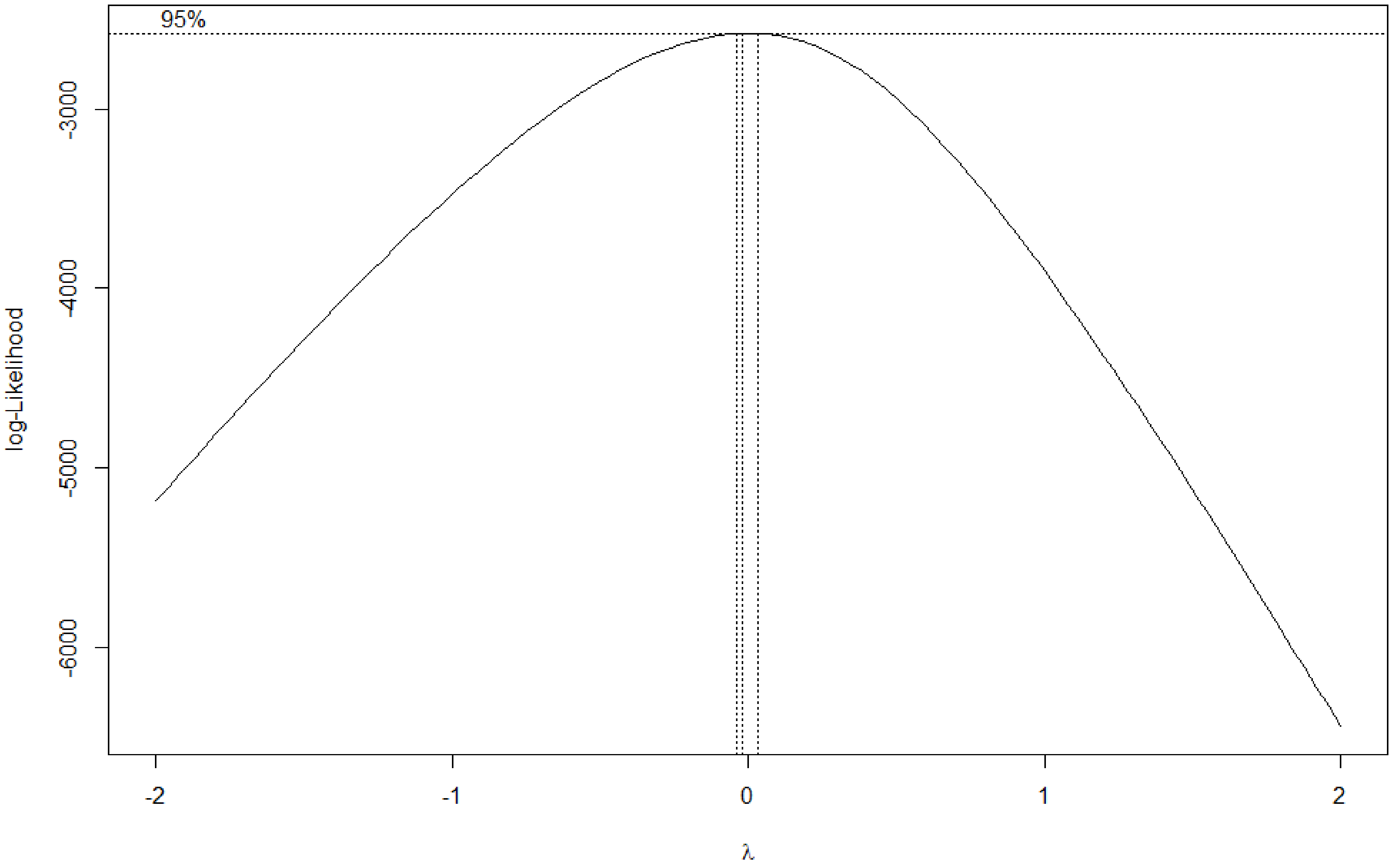
continent	location	date	new_cases	new_deaths	new_tests	new_vaccinations
Asia	India	12/20/2020	24337	333	1107681	
Asia	Indonesia	12/20/2020	6982	221	41914	
Asia	Iran	12/20/2020	6312	177		
Asia	Iraq	12/20/2020	1027	17		
Asia	Israel	12/20/2020	1874	25	73443	**7315**
Asia	Japan	12/20/2020	2455	35	12282	
Asia	Jordan	12/20/2020	2152	23	18536	
Asia	Kazakhstan	12/20/2020	681	1	21075	

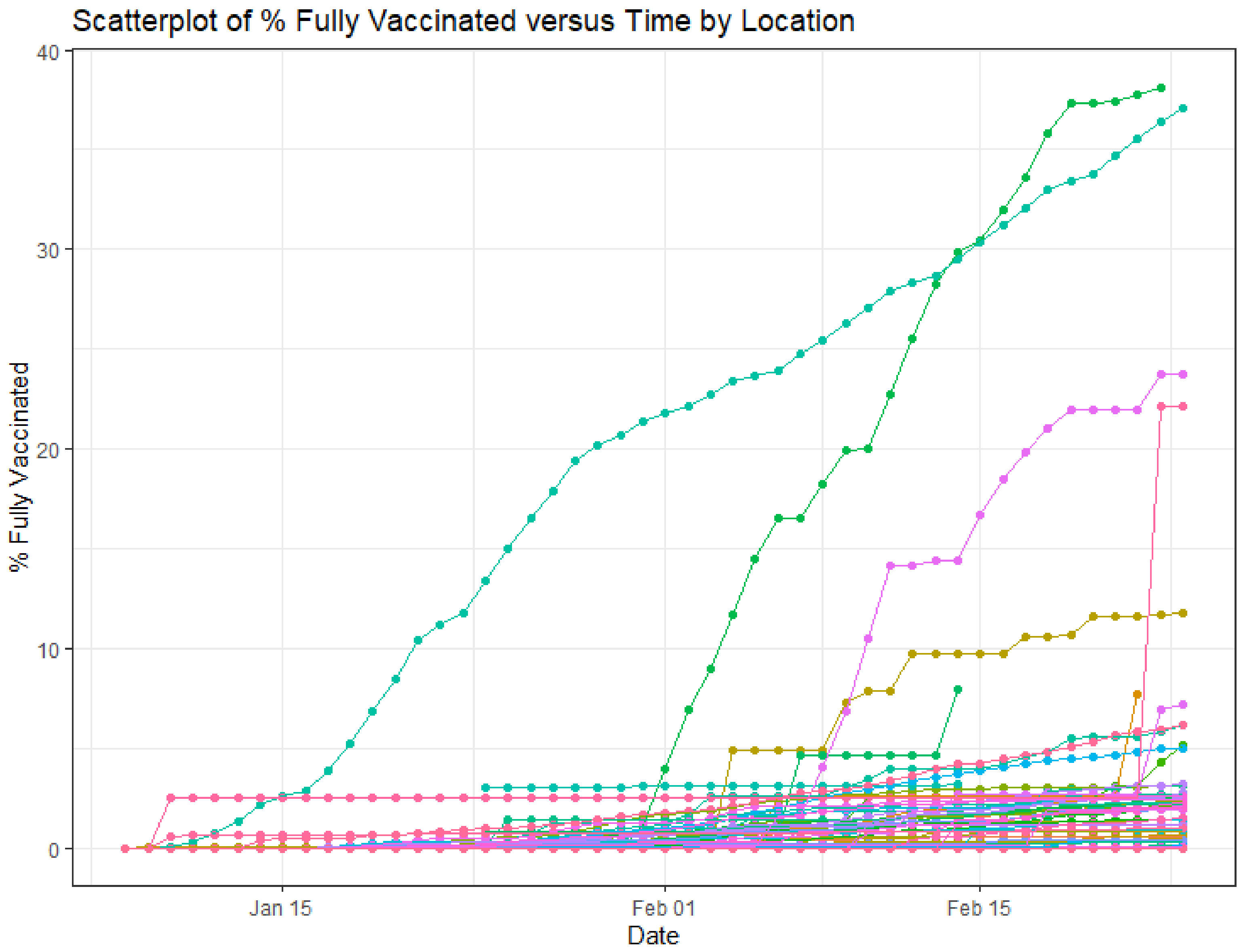
First recorded instance of vaccinations in dataset

$$y_i^{(\lambda)} = \begin{cases} \dfrac{y_i^\lambda - 1}{\lambda} & \text{if } \lambda \neq 0, \\ \ln(y_i) & \text{if } \lambda = 0, \end{cases}$$

https://medium.com/@ronakchhatbar/box-cox-transformation-cba8263c5206

Box-Cox Transformation for 20-day Lag – Lambda = 0 is within 95% confidence interval.




Scatterplot of % Fully Vaccinated versus Time by Location
% Fully Vaccinated
40
30
20
10
0
Jan 15
Feb 01
Feb 15
Date

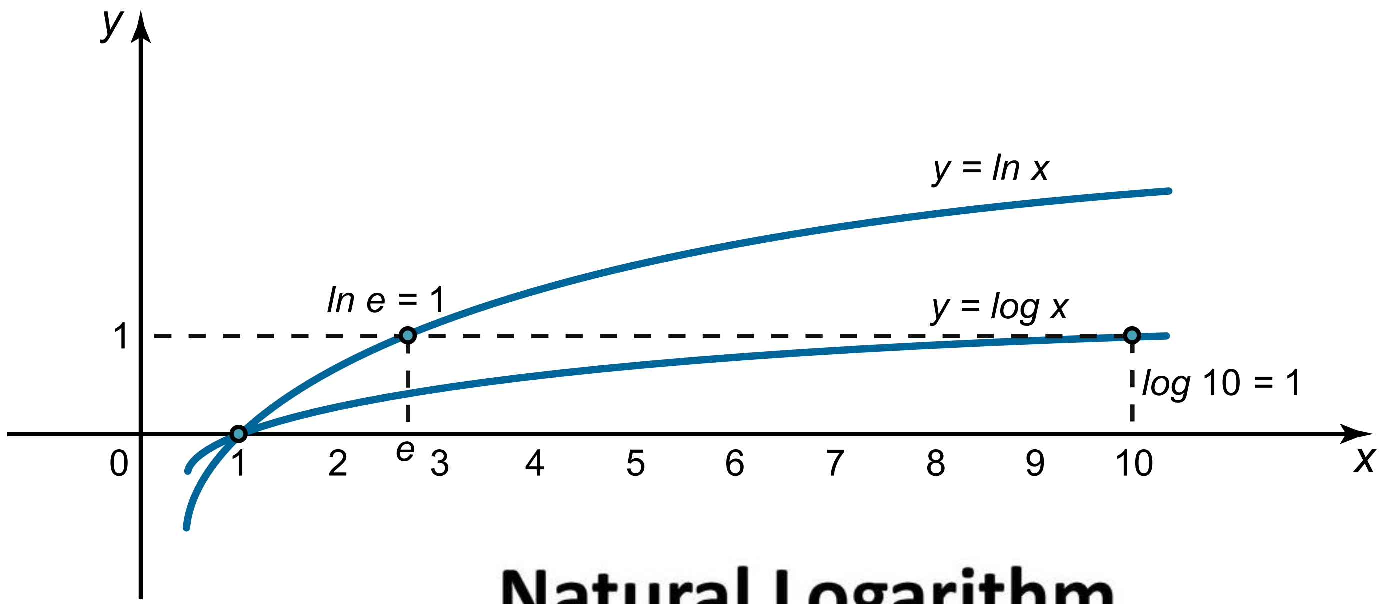
$$Residual\ standard\ error = \sqrt{\frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{\mathrm{df}}}$$

https://quantifyinghealth.com/residual-standard-deviation-error/

$$R^2 = 1 - \frac{SS_{residuals}}{SS_{total}}$$

$$\text{Adjusted R}^2 = 1 - \frac{SS_{residuals}/(n-K)}{SS_{total}/(n-1)}$$

https://www.graphpad.com/guides/prism/latest/curve-fitting/reg_interpreting_the_adjusted_r2.htm



https://www.math24.net/natural-logarithms/

Natural Logarithm

- **Natural Logarithm:** log with base e: $\log_e x$
  - **Notation:** ln x

https://slideplayer.com/slide/13798315/

$$\ln x = y \Leftrightarrow e^y = x$$

e is the base   *e* is the base

x is the argument   *y* is the exponent

$$\text{Regression Line} = \left[\left(\frac{1}{n-1}\sqrt{\frac{\sum(x-\mu_x)(y-\mu_y)}{\sigma_x\,\sigma_y}}\right)\left(\frac{\sigma_y}{\sigma_x}\right)\right]x + b$$

where $b = \bar{y} - m\bar{x}$

http://www.learningaboutelectronics.com/Articles/Regression-line-calculator.php

Example Code

```
#attach needed libraries
library(tidyverse)#DATA MANIPULATION
library(ggplot2)#PLOTS
library(dplyr)#DATA MANIPULATION
library(ggpubr)#PLOTS
library(car)#ANOVA
library(nortest)#NORMALITY TESTS
library(moments)#SKEWNESS
library(geoR)#BOXCOXTRANSFORMATIONS
library(gamlss)#checking fit of distributions
library(gamlss.dist)#checking fit of distributions
library(gamlss.add)#checking fit of distributions
library(lubridate)#for week aggregating
library(Hmisc)#lag function
library(caret)#RMSE and R2 functions
library(ggpmisc)#show equation and r2 on graph
library(graphics)#curved bestfit lines
```

```r
#create lag functions for new cases per million and people fully vaccinated
#21 day lag only
covid2 <- covid2 %>%
  group_by(location) %>%
  mutate(dif_new_cases_per_million = new_cases_per_million - Lag(new_cases_per_million, shift = 7),
         dif2_new_cases_per_million = new_cases_per_million - Lag(new_cases_per_million, shift = 14),
         dif3_new_cases_per_million = new_cases_per_million - Lag(new_cases_per_million, shift = 21),
         lag_people_fully_vaccinated_per_hundred = Lag(people_fully_vaccinated_per_hundred, shift = 7),
         lag2_people_fully_vaccinated_per_hundred = Lag(people_fully_vaccinated_per_hundred, shift = 14),
         lag3_people_fully_vaccinated_per_hundred = Lag(people_fully_vaccinated_per_hundred, shift = 21)
  )
```

```
#linear model
#dif3_new_cases_per_million vs lag3_people_fully_vaccinated_per_hundred daily
fit28 <- lm(dif3_new_cases_per_million ~ lag3_people_fully_vaccinated_per_hundred, data=covid20)
summary(fit28)
plot(covid20$lag3_people_fully_vaccinated_per_hundred,covid20$dif3_new_cases_per_million,
     xlab = "Fully vaccinated/100 three weeks prior",
     ylab = "Difference in new cases per million three weeks prior")
abline(lm(dif3_new_cases_per_million ~ lag3_people_fully_vaccinated_per_hundred, data=covid20))
#R^2 ADJ = 0.04
```

```r
for(i in 1:28){
    covidloop <- covid22 %>%
    group_by(location) %>%
    mutate(dif_new_cases_per_million = new_cases_per_million - Lag(new_cases_per_million, shift = i),
           lag_people_fully_vaccinated_per_hundred = Lag(people_fully_vaccinated_per_hundred, shift = i)
    )
    covidloop <- subset(covidloop, !is.na(lag_people_fully_vaccinated_per_hundred))
  covidloop <- subset(covidloop,lag_people_fully_vaccinated_per_hundred != 0)
    covidloop$log_lag_people_fully_vaccinated_per_hundred <- log(covidloop$lag_people_fully_vaccinated_per_hundred)
    covidloop$lag <- i
    fitloop <- lm(dif_new_cases_per_million ~ log_lag_people_fully_vaccinated_per_hundred, data=covidloop)
    plot(covidloop$log_lag_people_fully_vaccinated_per_hundred,covidloop$dif_new_cases_per_million,
       xlab = paste("Natural Log of Fully vaccinated/100 ",i," days prior"),
       ylab = paste("Difference in new cases per million ",i," days prior"))
  abline(lm(dif_new_cases_per_million ~ log_lag_people_fully_vaccinated_per_hundred, data=covidloop))
    intercept[i] <- fitloop$coefficients[1]
  vaccine_effect_coef[i] <- fitloop$coefficients[2]
  rsquared[i] <- summary(fitloop)$r.squared
  adjrsquared[i] <- summary(fitloop)$adj.r.squared
    if(i==1){
    covid_combined_lag <- covidloop
  } else {
    covid_combined_lag <- rbind(covid_combined_lag, covidloop)
  }
    assign(paste("covid_loop_",i,sep=""), covidloop)
  assign(paste("fit_loop_",i,sep=""), fitloop)
}
```

```r
library(ggplot2)
library(gganimate)
theme_set(theme_bw())
library(transformr)
covid_combined_lag_2 <- subset(covid_combined_lag, lag > 6)
p <- ggplot(
  covid_combined_lag_2,
  aes(x = log_lag_people_fully_vaccinated_per_hundred, y=dif_new_cases_per_million)
) +
  geom_point() +
  scale_x_continuous(limits = c(-4.60518,3.555),
                     breaks = c(-4.60518, -2.303, 0, 2.303, 3.555),
                     label=c("0.01 %","0.1 %","1 %","10 %","35 %")) +
  labs(x = "Percent Vaccinated (Log Scale)", y = "Difference in New Cases per Million") +
  geom_smooth(method='lm', formula= y~x)
p
anim <- p +
  transition_states(lag,
                    transition_length = 1,
                    state_length = 4,
                    nframes = 140)+
  ggtitle('{closest_state} Day Difference')
anim
```

```
#from cdc website: It typically takes a few weeks for the body to build immunity
#(protection against the virus that causes COVID-19) after vaccination.
#https://www.cdc.gov/coronavirus/2019-ncov/vaccines/facts.html
```