UCDAVIS
COLLEGE OF LETTERS AND SCIENCE

Team 1
Yunhao Yang
Liuyi Bao
Jun Wei
Xiaoyi Xu

Analysis of the global pandemic—concerning the cases and deaths

Outline

Introduction

Data Cleaning

Descriptive analysis

Inferential analysis

Regression model

Causal interpretation

Questions to address

Summary

Introduction

About Dataset: 100725 observations and 8 different variables.

Quantitative variables：New_cases, New_deaths, Cumulative_cases,Cumulative_deaths

Categorical variables: Country,Country_code,WHO_region, Date_reported

Questions of interest

1. How the number of new cases and new deaths change with the time in each region？
2. Whether there is any differences in new cases and new deaths each month between different regions?
3. How can we describe the relationship between new cases and new deaths?

Data Cleaning

Data: Create monthly dataset with sum of each quantitative variables, filter the Date_reported in March.

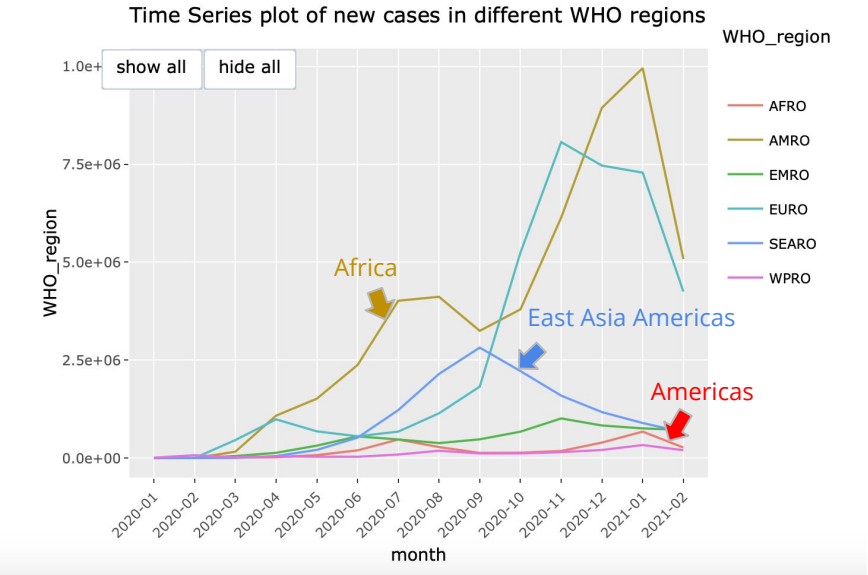
month	Country_code	Country	WHO_region	New_cases	Cumulative_cases	New_deaths	Cumulative_deaths
2020-01	AF	Afghanistan	EMRO	0	0	0	0
2020-02	AF	Afghanistan	EMRO	5	30	0	0
2020-03	AF	Afghanistan	EMRO	161	1141	4	22
2020-04	AF	Afghanistan	EMRO	2005	26299	60	837
2020-05	AF	Afghanistan	EMRO	13009	224580	190	4962
2020-06	AF	Afghanistan	EMRO	16265	750486	485	14926

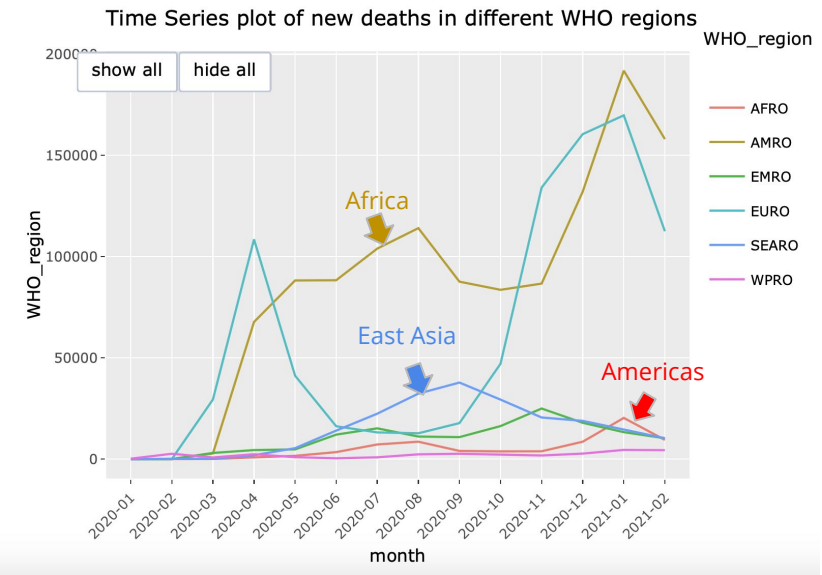
Data_region: The monthly dataset are cutted in region.

month <chr>	Region <fctr>	RNew_cases <dbl>	RNew_deaths <dbl>
2020-11	AFRO	180789	3872
2020-12	AFRO	395399	8584
2021-01	AFRO	675007	20347
2021-02	AFRO	269734	9487
2020-01	AMRO	19	0
2020-02	AMRO	81	0

Descriptive analysis

Q1: How the number of new cases and new deaths change with the time in each region？

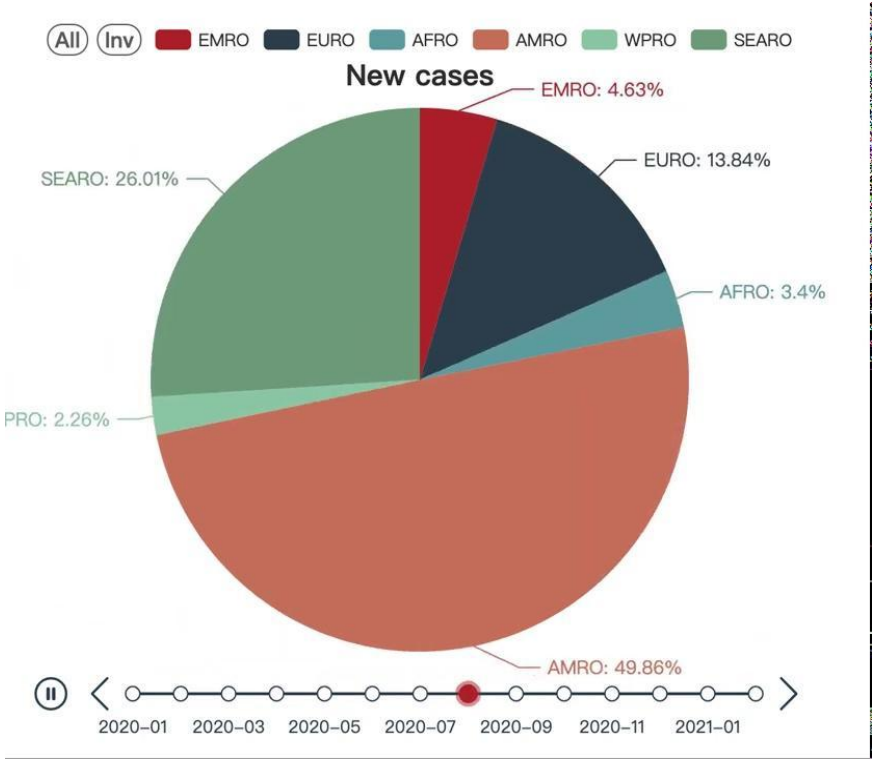


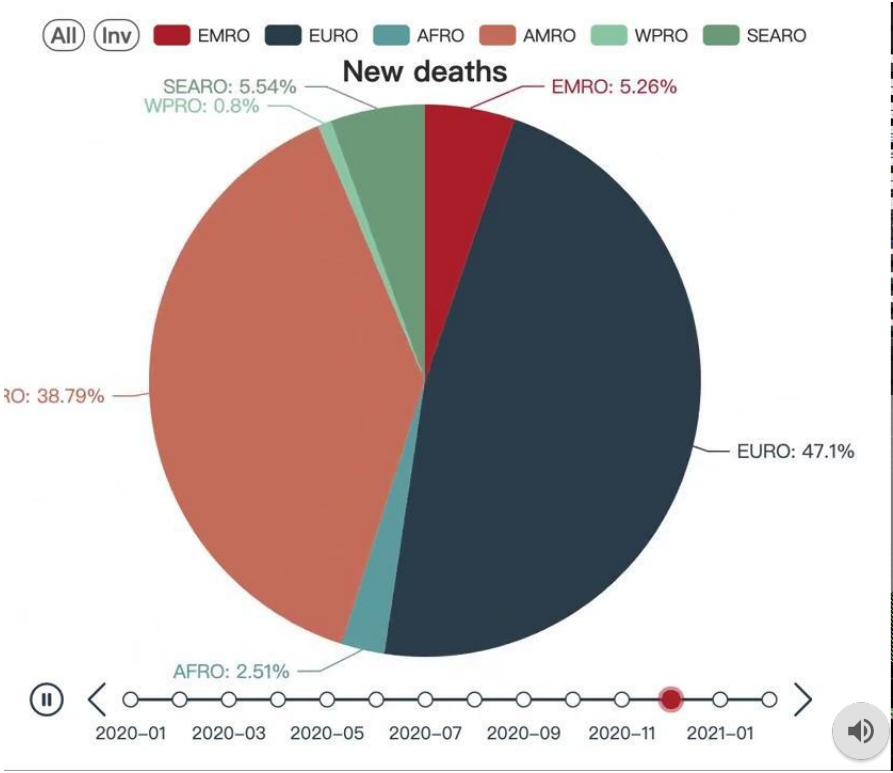


Similar trend

Descriptive analysis

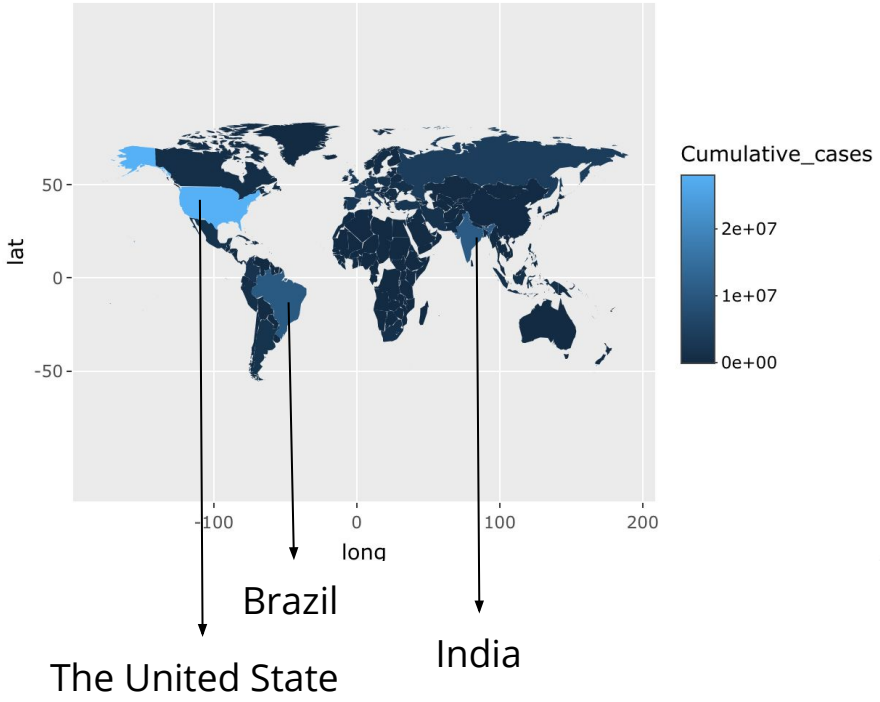
Q2: Whether there is any differences in new cases and new deaths each month between different regions?



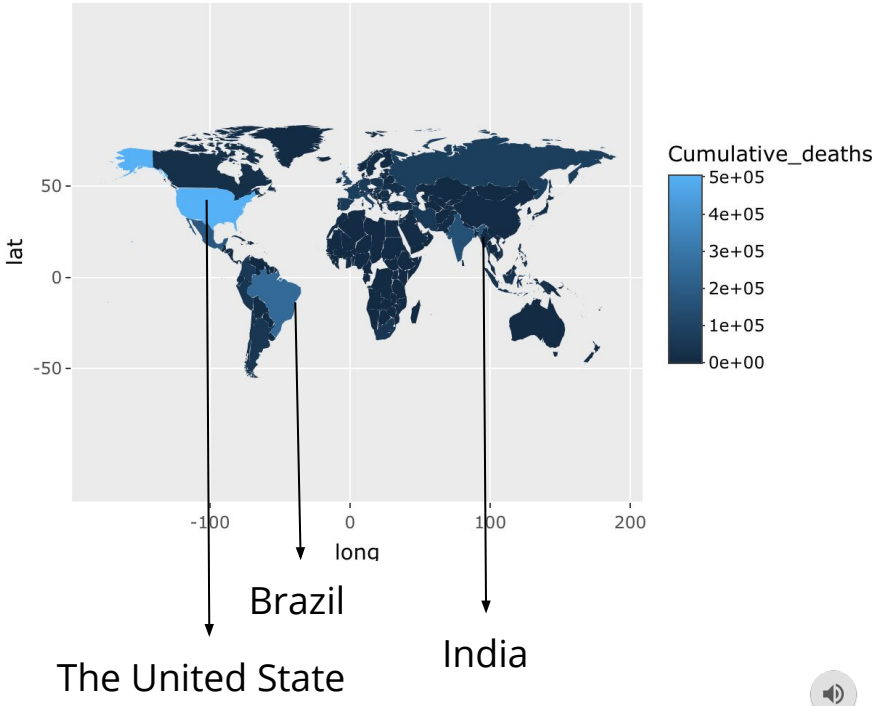


Descriptive analysis

The number of cumulative cases over countries

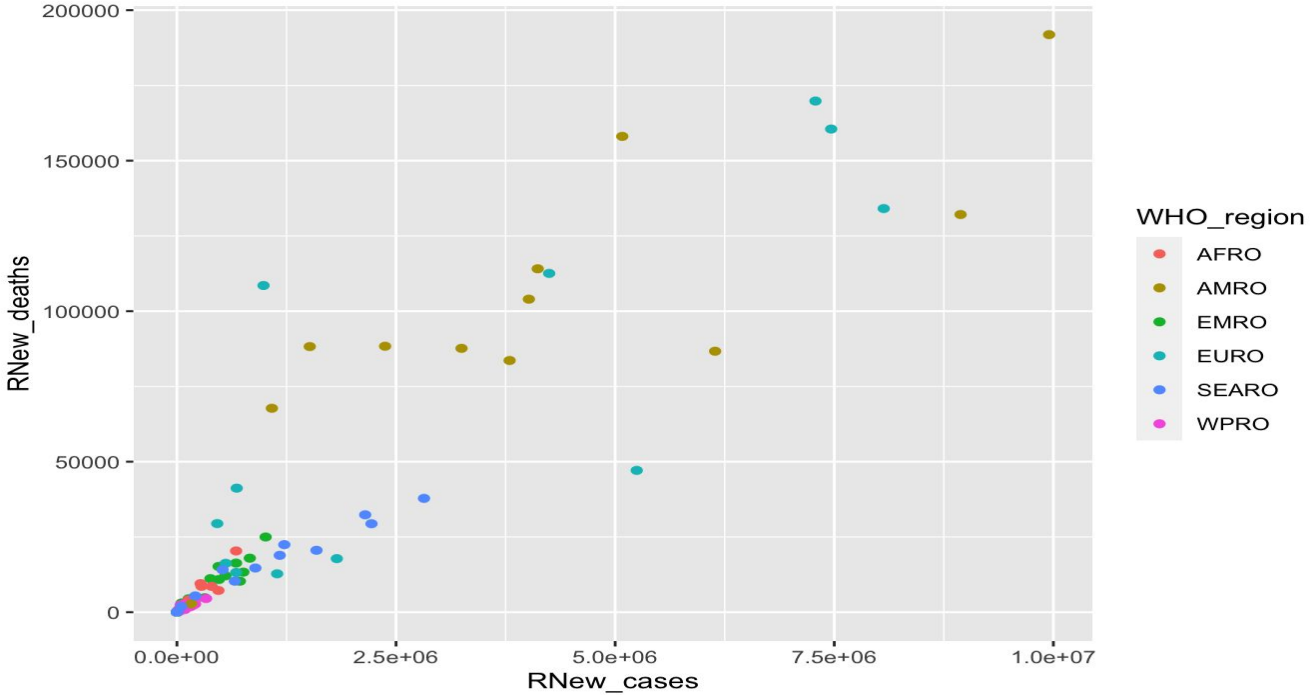


The number of cumulative deaths over countries



Inferential analysis—— Regression Model

Q3: How to describe the relationship between new cases and new deaths



Inferential analysis-- Regression Model

```
## Call:
## lm(formula = RNew_deaths ~ RNew_cases:factor(WHO_region), data = data_region)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -57112  -7490  -3196   1872  82194
##
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     8.261e+03  3.231e+03   2.557   0.0125 *
## RNew_cases:factor(WHO_region)AFRO  4.206e-03  2.028e-02   0.207   0.8363
## RNew_cases:factor(WHO_region)AMRO  1.909e-02  1.222e-03  15.626   <2e-16 ***
## RNew_cases:factor(WHO_region)EMRO  9.700e-03  1.053e-02   0.921   0.3598
## RNew_cases:factor(WHO_region)EURO  1.829e-02  1.402e-03  13.051   <2e-16 ***
## RNew_cases:factor(WHO_region)SEARO  9.838e-03  4.310e-03   2.282   0.0252 *
## RNew_cases:factor(WHO_region)WPRO -2.924e-02  3.969e-02  -0.737   0.4635
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 19400 on 77 degrees of freedom
## Multiple R-squared:  0.8398, Adjusted R-squared:  0.8273
## F-statistic: 67.25 on 6 and 77 DF,  p-value: < 2.2e-16
```

```
## Call:
## lm(formula = RNew_deaths ~ RNew_cases:factor(WHO_region) - 1,
##     data = data_region)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -56243   -540    252   2633  89062
##
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)
## RNew_cases:factor(WHO_region)AFRO  0.025519   0.019135   1.334 0.186213
## RNew_cases:factor(WHO_region)AMRO  0.020437   0.001141  17.910  < 2e-16 ***
## RNew_cases:factor(WHO_region)EMRO  0.021999   0.009692   2.270 0.025986 *
## RNew_cases:factor(WHO_region)EURO  0.019704   0.001334  14.772  < 2e-16 ***
## RNew_cases:factor(WHO_region)SEARO 0.014404   0.004060   3.548 0.000661 ***
## RNew_cases:factor(WHO_region)WPRO  0.015761   0.036810   0.428 0.669701
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 20080 on 78 degrees of freedom
## Multiple R-squared:  0.8775, Adjusted R-squared:  0.8681
## F-statistic: 93.11 on 6 and 78 DF,  p-value: < 2.2e-16
```

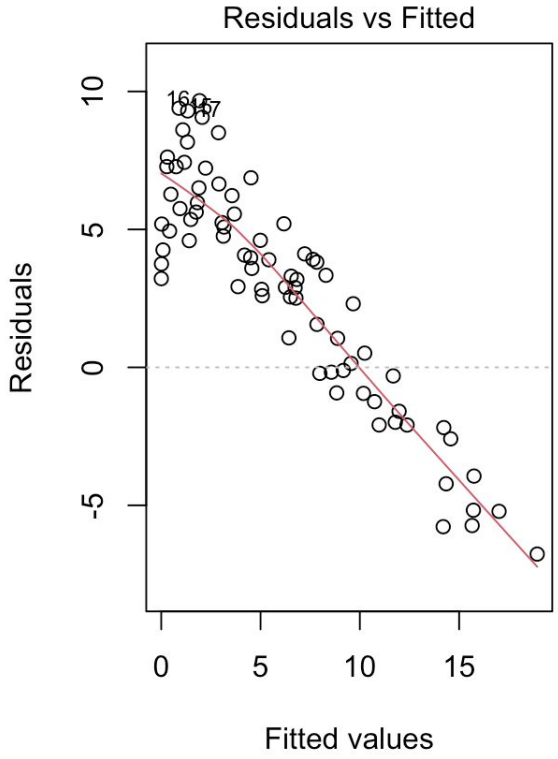
Inferential analysis–– Regression Model

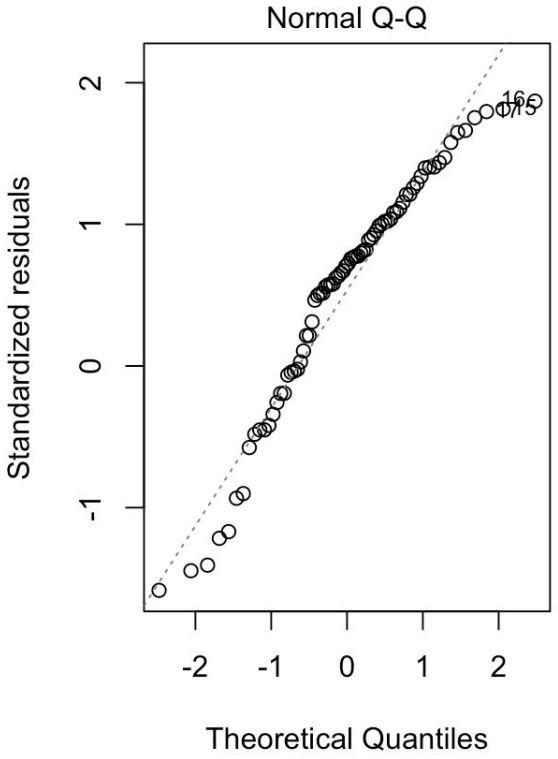
model form:

$$Z_j = \beta_1 Y_j R_1 + \beta_2 Y_j R_2 + \beta_4 Y_j R_3 + \beta_4 Y_j R_4 + \beta_5 Y_j R_5 + \beta_6 Y_j R_6 + \epsilon_j, \epsilon_j \sim N(0, \sigma^2), i.i.d$$

Z_j stands for the monthly new deaths, Y_j stands for the monthly new cases and $R_1, R_2, \ldots R_6$ are the dummy variables representing different WHO regions.

Inferential analysis-- Diagnosis





Inferential analysis—— Regression Model

Final model:

$$logZ_j = \beta_1 Y_j R_1 + \beta_2 Y_j R_2 + \beta_4 Y_j R_3 + \beta_4 Y_j R_4 + \beta_5 Y_j R_5 + \beta_6 Y_j R_6 + \epsilon_j,\ Z_j > 0$$

Inferential analysis-- Regression Model

```
## Call:
## lm(formula = RNew_deaths ~ RNew_cases:factor(WHO_region) - 1,
##     data = data_change)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.7476 -0.1285  3.6686  5.6639  9.6655
##
## Coefficients:
##                                      Estimate Std. Error t value Pr(>|t|)
## RNew_cases:factor(WHO_region)AFRO  2.319e-05  4.937e-06   4.697 1.28e-05 ***
## RNew_cases:factor(WHO_region)AMRO  1.903e-06  2.944e-07   6.463 1.17e-08 ***
## RNew_cases:factor(WHO_region)EMRO  1.417e-05  2.501e-06   5.666 3.01e-07 ***
## RNew_cases:factor(WHO_region)EURO  1.952e-06  3.442e-07   5.673 2.93e-07 ***
## RNew_cases:factor(WHO_region)SEARO 5.576e-06  1.047e-06   5.323 1.17e-06 ***
## RNew_cases:factor(WHO_region)WPRO  4.272e-05  9.497e-06   4.498 2.66e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.18 on 70 degrees of freedom
## Multiple R-squared:  0.7162, Adjusted R-squared:  0.6919
## F-statistic: 29.45 on 6 and 70 DF,  p-value: < 2.2e-16
```

Causal interpretation

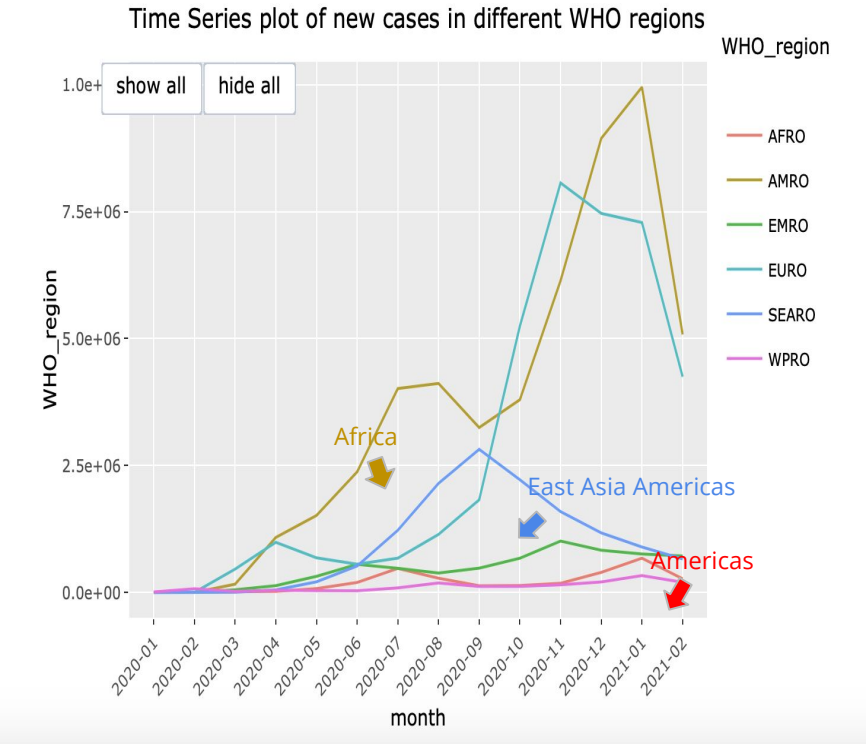
Unable to draw the conclusion about causal relationships

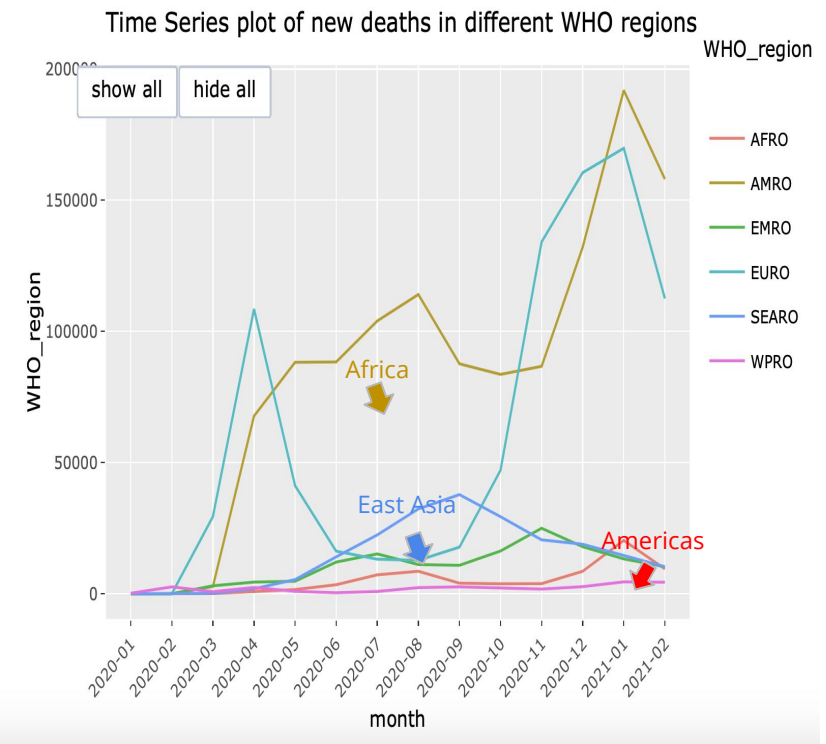
Association: monthly new cases and monthly new deaths(same trend but different in different regions)

Questions to address

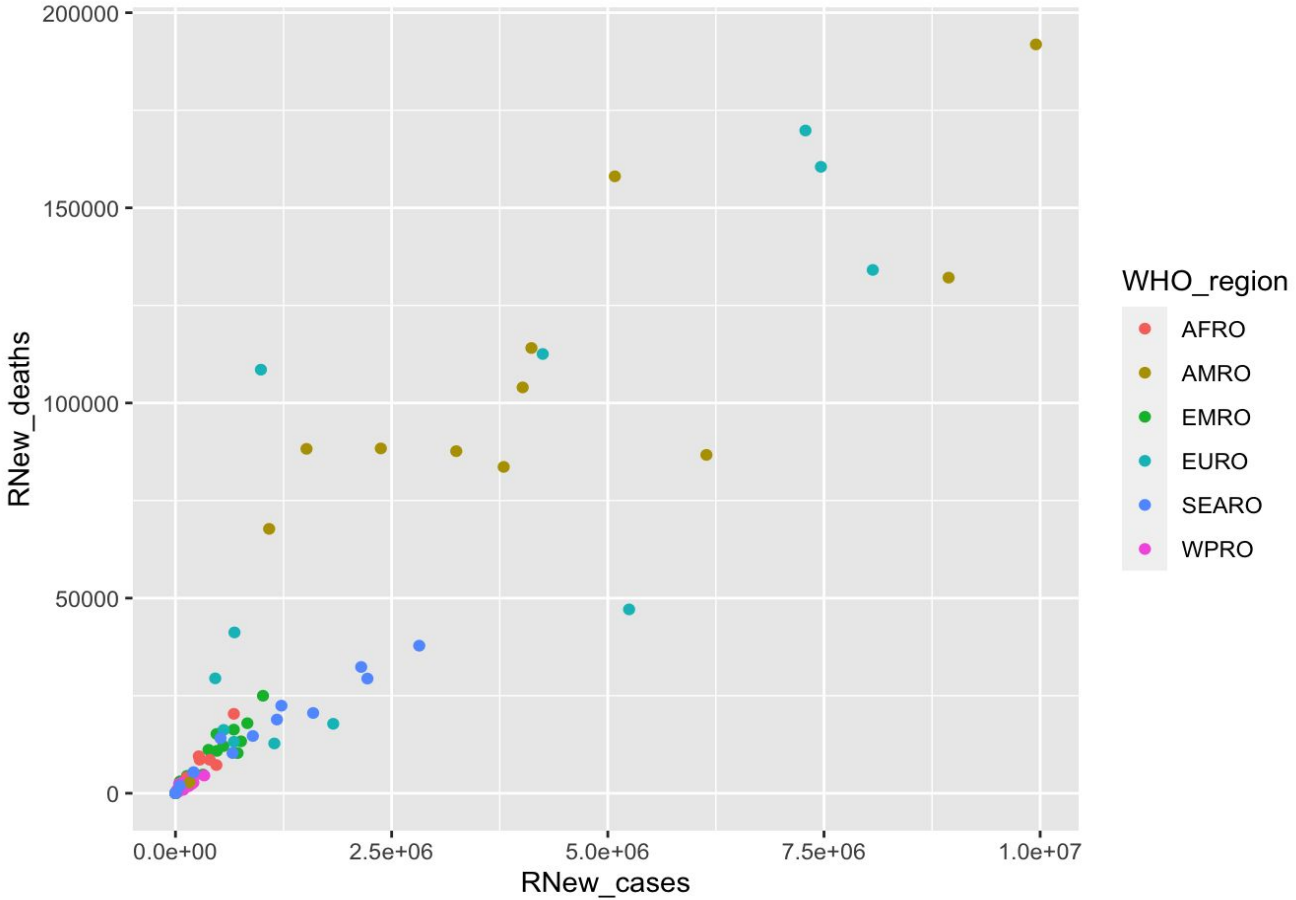
Why do we use the monthly data over regions?

Why do we drop the variables about time in linear model?





Similar trend


200000
150000
100000
50000
0
RNew_deaths
0.0e+00
2.5e+06
5.0e+06
7.5e+06
1.0e+07
RNew_cases
WHO_region
AFRO
AMRO
EMRO
EURO
SEARO
WPRO

Questions to address

How can we explain the association between new deaths and new cases?

Why can’t we draw the conclusion about casual relationships?

How can we improve our model or report?

Summary

Achievements

--project

1. Question 1: visualization plots
2. Question 2: ANOVA model
3. Question 3: multiple regression model

--presentation

Shortcomings and improvements

1. Basic assumptions
2. More data and more variables