Data Analysis: COVID-19 Worldwide between 2020 and 2021

Team 9: Nathan Castle, Shanshan Chen and Fangyi Wang

Introduction: Understanding the Impact of COVID-19

- COVID-19 Global Pandemic
  - WHO declared COVID-19 a global pandemic in March 2020
  - Months prior, COVID-19 had never been detected in humans before

- World Health Organization, or WHO
  - Works within the United Nations to help coordinate international health
  - Data collected from censuses, healthcare systems, etc.

- Combating COVID-19
  - Lockdown orders, social distancing, facial coverings
  - Different policies for different regions and countries

Q1: Why did this analysis only involve WHO's COVID-19 dataset and not other sources?

Potential Questions and Hypotheses

1. Are the different World Health Organization regions associated with different daily average number of new cases and deaths from COVID-19?

H1: We predict WHO regions will be associated with average new daily cases and deaths of COVID-19

2. If so, is there a WHO region that's associated with the highest average daily number of new cases and deaths of COVID-19?

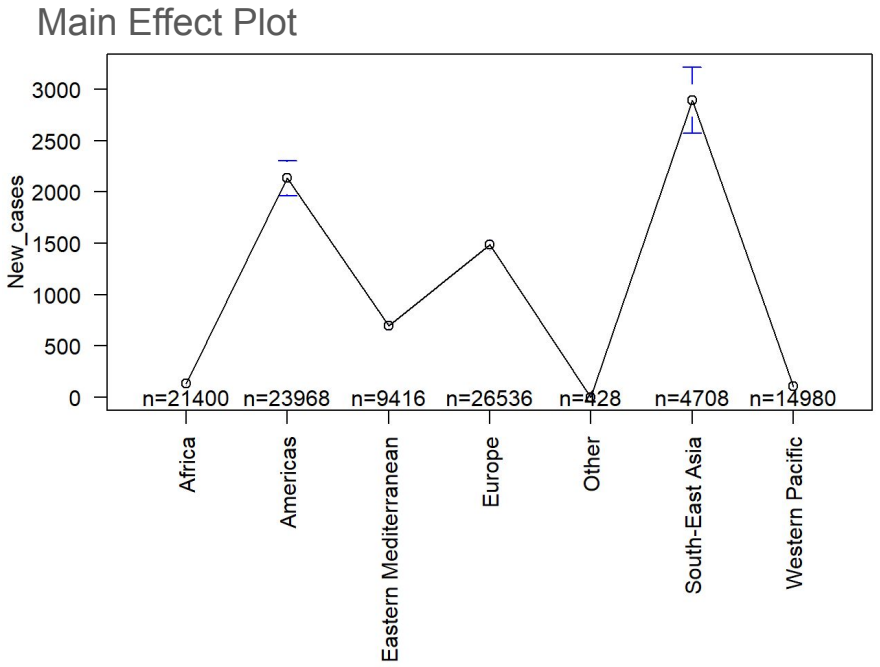
H2: We predict a highly populated WHO region will be associated with highest average daily new cases and deaths of COVID-19.

Descriptive Analysis

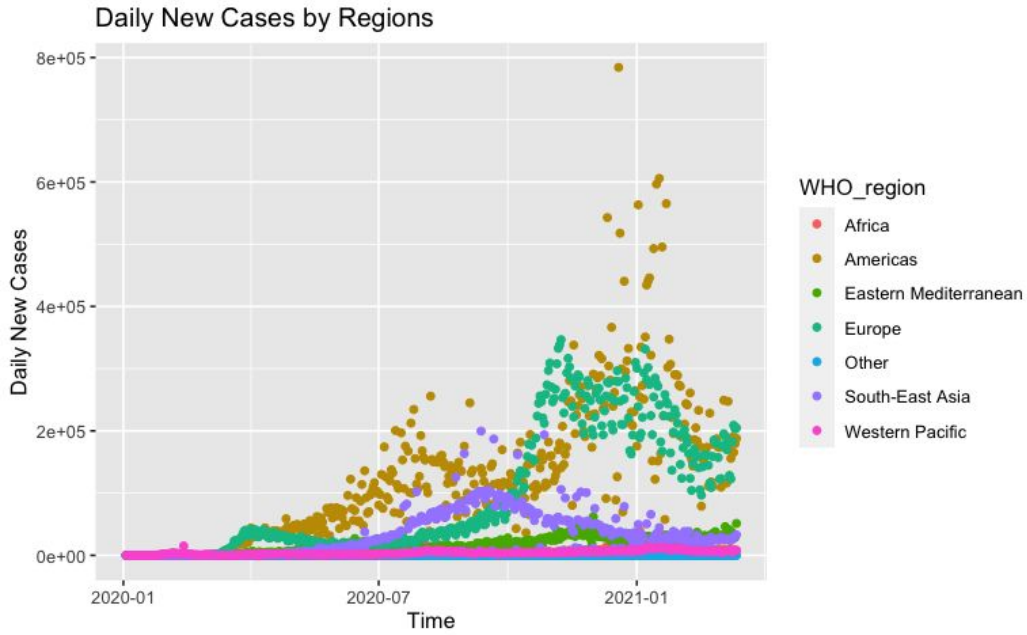
- The dataset included over 100 thousand observations from January 3rd 2020 to March 5th 2021.
- Each observation was composed of date, country, region, the daily new cases, daily cumulative cases, daily new death, and daily cumulative death.
- 856 observations with missing values, and 100 observations with negative values.
- Removed missing values and negative values observations.

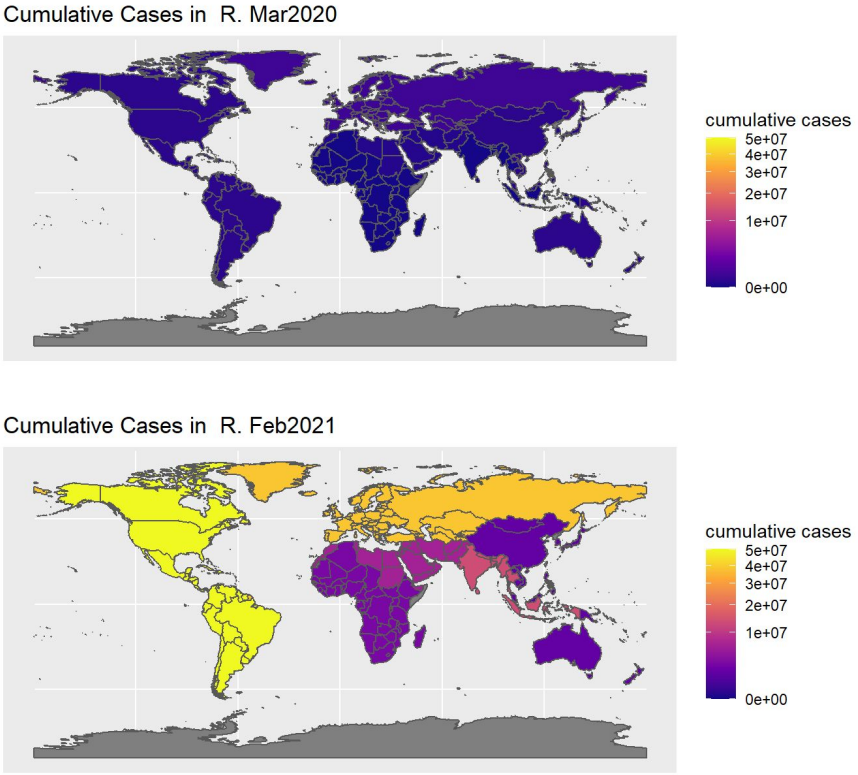
Descriptive Analysis

	Daily new cases	Daily new death
Mean	1147	25
Max	402270	6409

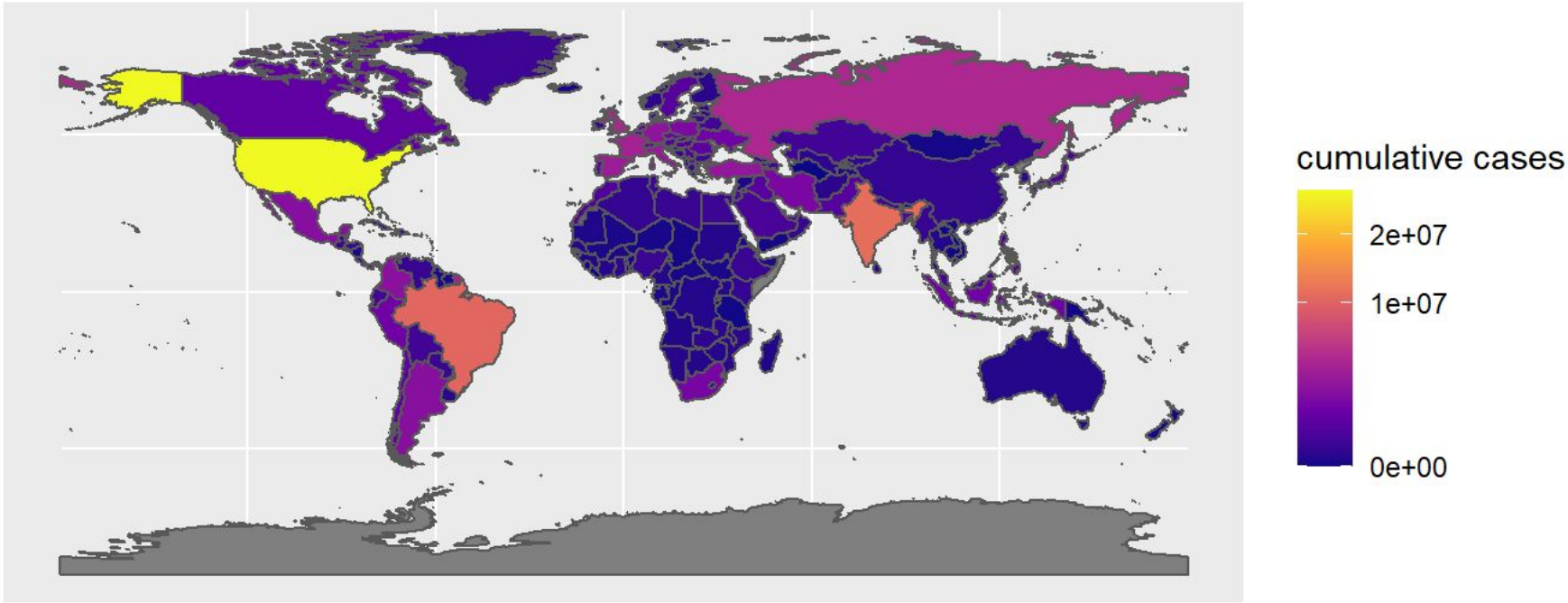


Daily new cases & World maps of cumulative cases





World maps of cumulative cases by country



Q2: Why did you include world maps of monthly cumulative cases by country and by region? What information can they provide in addition to your longitudinal plot?

Proposed Model: Mixed Effect Model

$$Y_{ijk} = \mu_{..} + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ijk},\ k = 1, \ldots, n_{ij}, j = 1, \ldots, b, i = 1, \ldots, a$$

Model Explanation:

- Alpha i are fixed effects of Region
- Beta j are random effects of Month to reduce the variance

Strong Assumption:

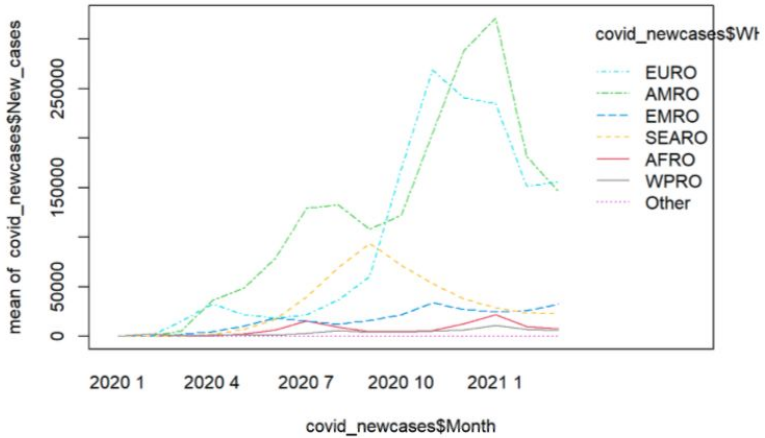
- Beta j are i.i.d normal distribution (Month is newly introduced covariate variable)

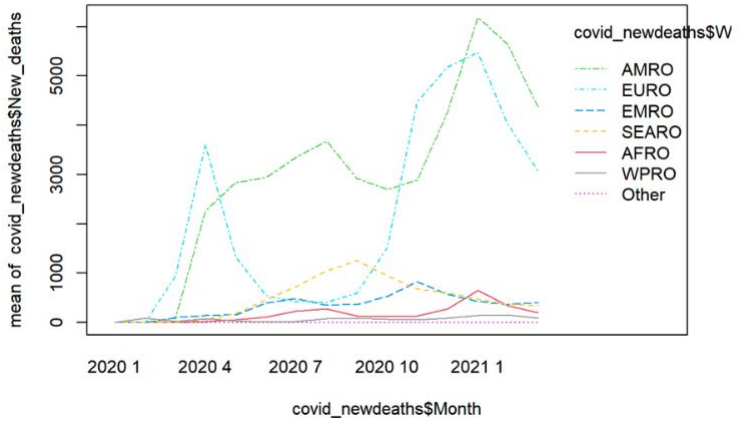
Transformation Data

- **Sum up** daily new cases/death of all countries in each region
- Introduce new variable **Month** from Date_reported

Q3:Why did you choose to use a mixed-effects model for analyzing this particular COVID-19 dataset?

Interaction Plots





- Unparallel lines indicates interaction term are involved.

Fit the proposed model

	New Cases	New Deaths
Random Effects	Std.Dev	Std.Dev
Month	26018	435.4
Month:Region	47078	943.7
Residual	16316	498.1
Fixed Effects		
Intercept(AFRO)	6732	172.0
AMRO	**113547**	**2767.7**
EMRO	9512	172.7
EURO	88487	1932.5
Other	-6718	-171.4
SEARO	24405	307.0
WPRO	-2781	-101.3

F test for the interested question

Null Hypothesis $H_0 : \alpha_1 = \alpha_2 = \cdots = \alpha_7$

Alternative Hypothesis H_a : not all α_i are the same.

Reject Rule: Reject Null Hypothesis at level 0.05 if the p-Value less than 0.05

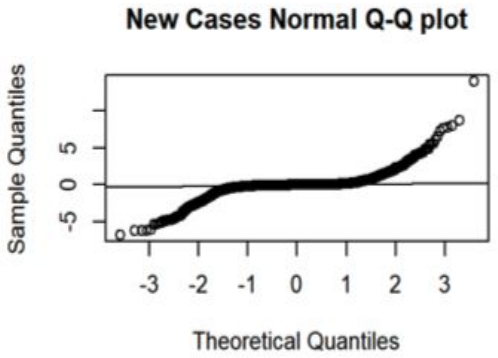
Statistics F: p-value of new cases 7.927e-13
p-value of new deaths 2.2e-16

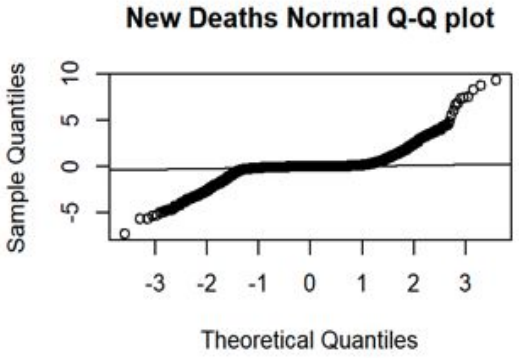
Conclusion: the region effects are significant.

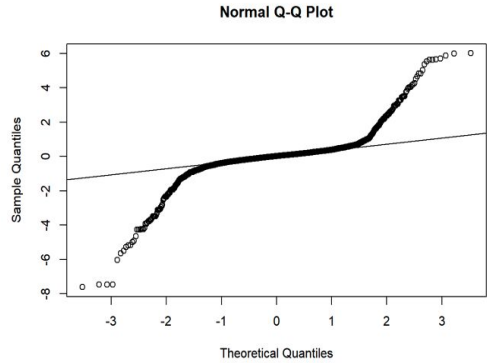
Sensitivity Analysis

Q4: Why did your team choose to stay with the original model instead of choosing to use the log-transformation model after doing the transformation and comparing their diagnostic plots?

After log transformation







Conclusions and Caveats

From the analysis, questions of interest can be answered:

1. WHO regions associated with average new daily COVID-19 case and death numbers
2. Region of the Americas associated with the highest average number of daily new cases and deaths

There are several possible concerns of the analysis:

- Model assumption violations, no “good” transformation identified
- Dataset contained possible errors in recorded values
- Observational study, unknown data collection specifics, and potential outcome assumptions violated ⇒ difficult to ascertain causal inference

Questions for our project

1. Why did your team choose to only analyzing the COVID-19 dataset from WHO versus combining multiple similar COVID-19 datasets?
2. Why did you include world maps of monthly cumulative cases by country and by region? What information can they provide in addition to your longitudinal plot?
3. Why did your team choose to stay with the original model instead of choosing to use the log-transformation model after doing the transformation and comparing their diagnostic plots?
4. Why did you choose to use a mixed-effects model for analyzing this particular COVID-19 dataset?
5. The covid cases data are essentially counts data, and a good way to analyze count data is using Poisson regression (log transformation). However, your log-transformed model didn't seem to improve compared to the original model, what could be the reason for that?

Citations

- About WHO. Retrieved March 05, 2021, from https://www.who.int/about

- CDC COVID Data Tracker. (2021). Retrieved March 05, 2021, from https://covid.cdc.gov/covid-data-tracker/#trends_dailytrendscases

- WHO | Health Statistics. (2014, May 15). Retrieved March 05, 2021, from https://www.who.int/features/qa/73/en/