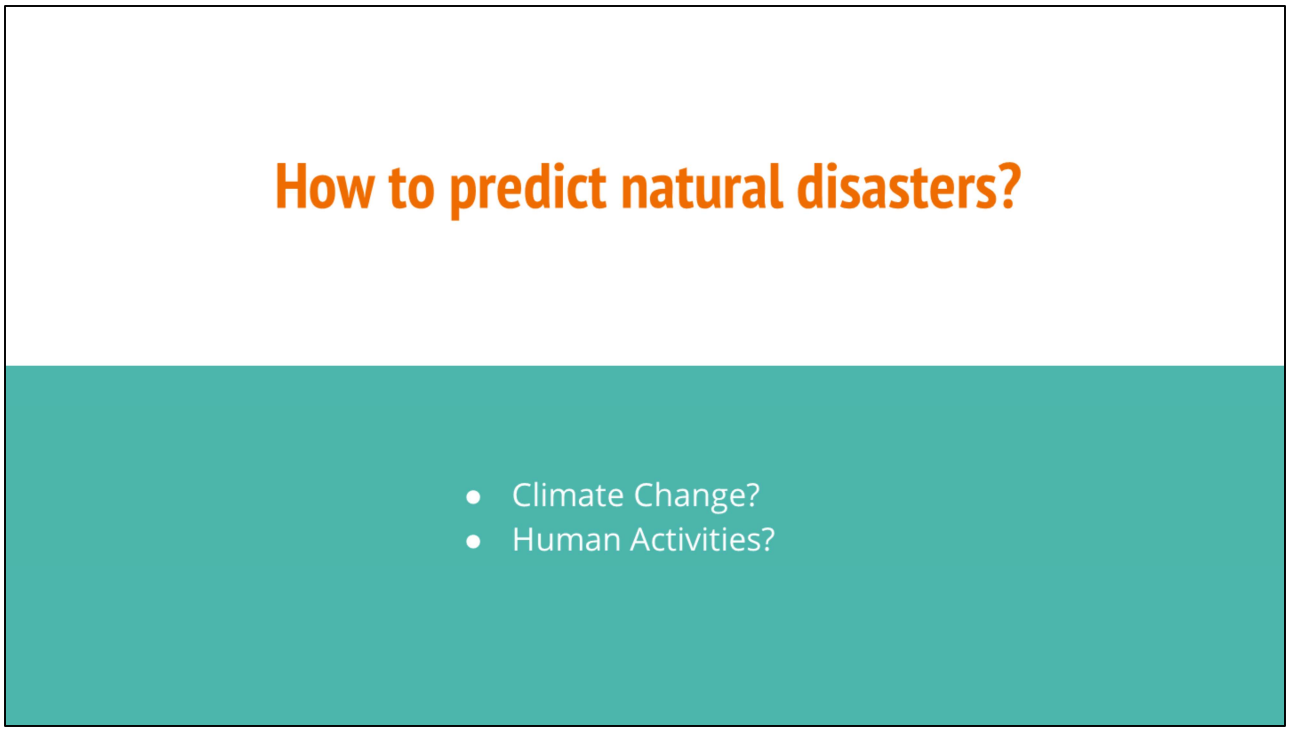
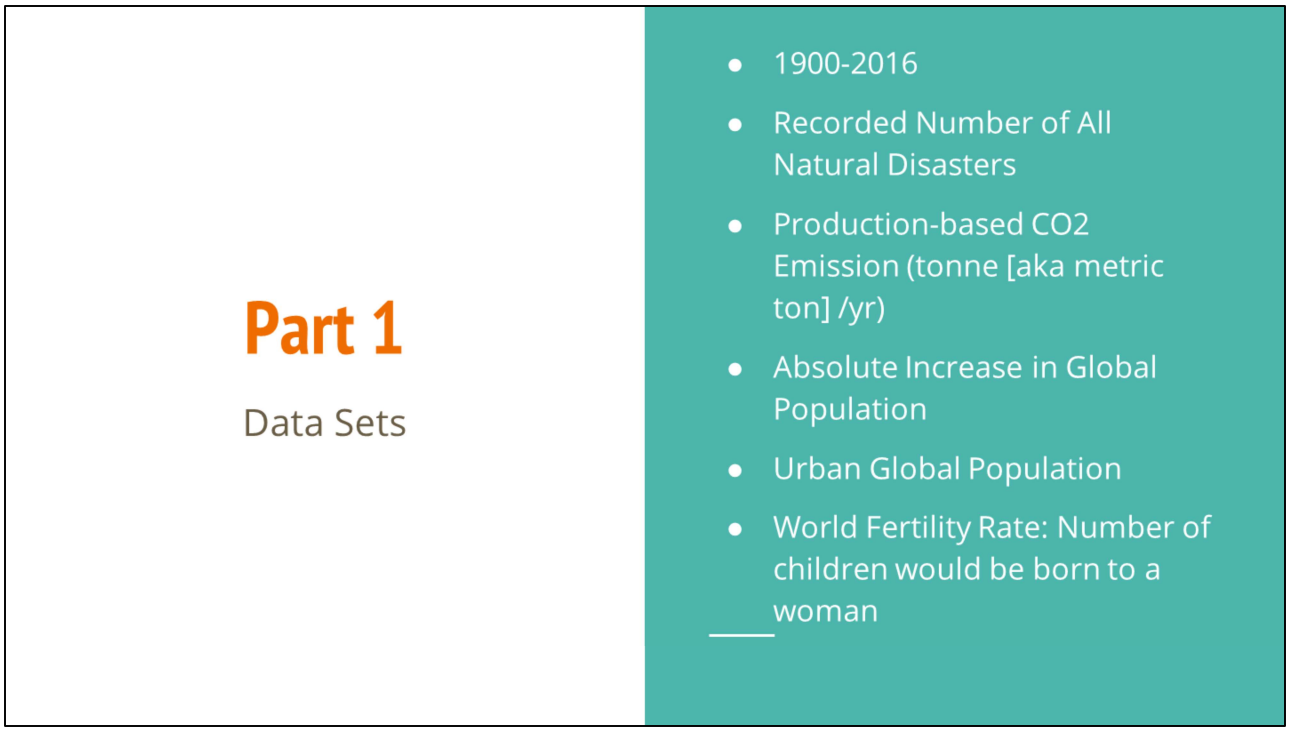
Predicting Natural Disasters

Sasha Bakker, Ka Wing Cheung,
and Vishvesh Gandhi



For this project, we would like to find a way to predict the number of natural disasters. Studies have shown that climate change is heavily influenced by human activities and contributes to some common disasters, such as flood and drought. Thus, we are considering predictors that are known to influence climate change, such as CO2 emissions and urbanization.



Our data set on global recorded natural disasters includes 10 natural disasters, such as floods, droughts, and wildfires.

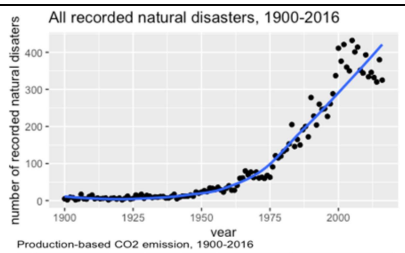
Our covariates are Production-based CO2 Emission, Absolute Increase in Global Population, Urban Global Population, and the Number of children that would be born to a woman.

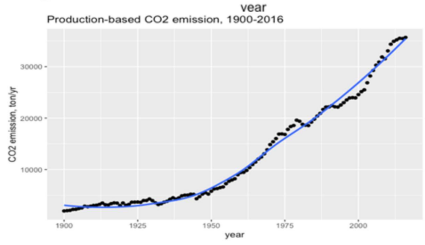
All the data sets come from ourworldindata.org, which collected data from other organizations and papers.

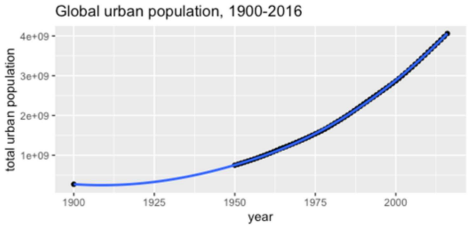
In the next slide we have plotted each variable

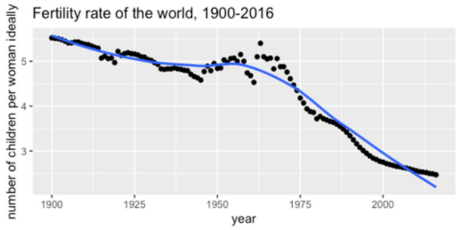
against time from 1900-2016. Each of the variables has a general trend.

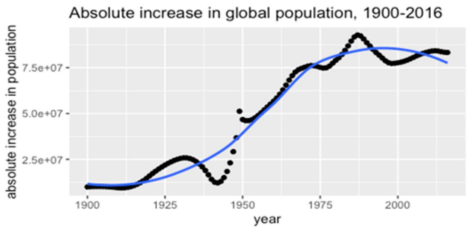
Scatter Plots of our variables by time (1900-2016)







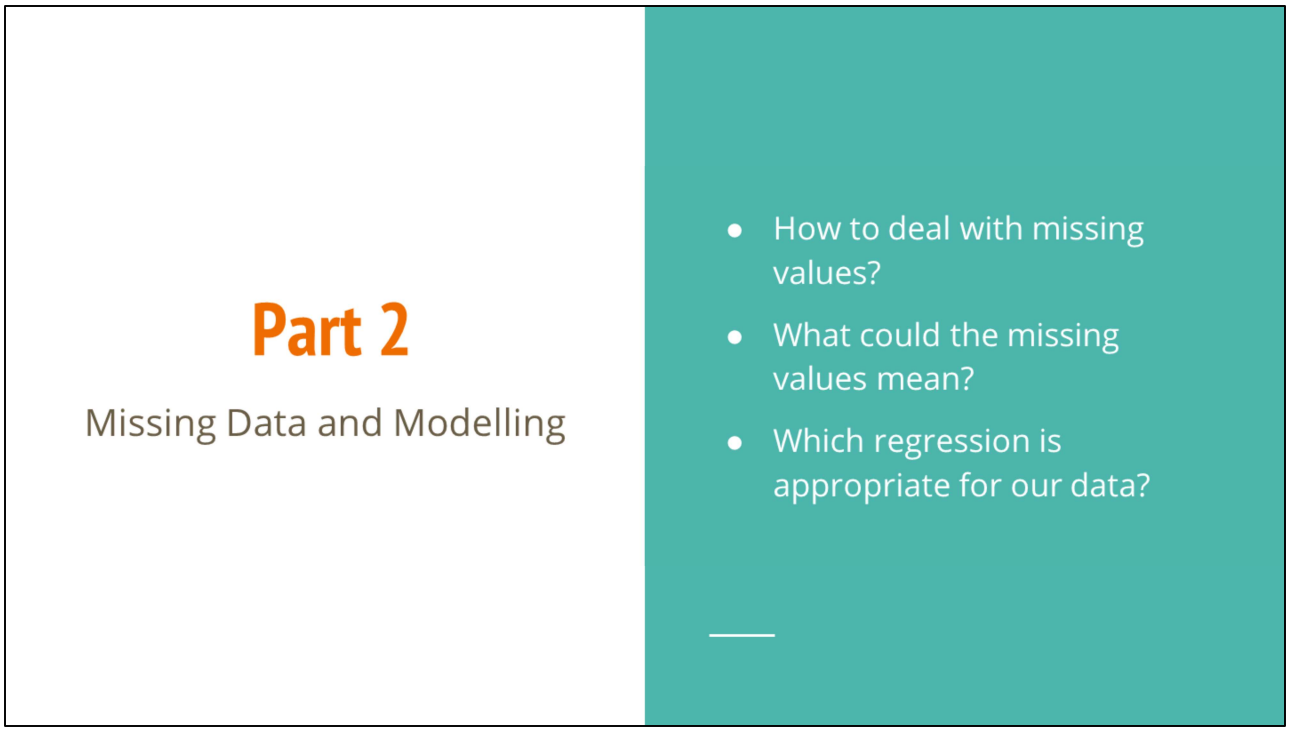




A Subset of Our Data

year <int>	co2 <dbl>	all.num <int>	abs.inc.pop <int>	fertility.rate <dbl>	urban.pop <dbl>
1969	13846.43	62	74622178	4.88	1319832960
1970	14845.81	77	75246977	4.88	1354215040
1971	15409.72	61	75860226	4.76	1388834048
1972	16023.99	63	76127226	4.61	1424734976
1973	16893.36	60	76011655	4.47	1462178048
1974	16923.58	68	75665453	4.35	1501134976
1975	16825.64	63	75142842	4.18	1538625024
1976	17800.97	91	74780357	4.07	1577376000

After filtering and rearranging variables, we combined the separate data sets into one. And here is a subset of our data.



Our data set has some missing values. So one of our goals is to deal with them and to make sense of them.

How to deal with the missing values?

Drought	46 years, 39.3%
Earthquake	4 years, 3.42%
Extreme Temperature	62 years, 53.0%*
Extreme Weather	4 years, 3.4%
Flood	24 years, 20.5%
Landslide	39 years, 33.3%
Mass Movement (dry) - any type of downslope movement of earth materials	85 years, 72.65%*
Impact - a type of extraterrestrial hazard caused by the collision of the Earth with a meteoroid, asteroid or comet.	116 years, 99.1%*
Volcanic Activities	41 years, 35.0%
Wildfire - can be triggered by lightning or human actions	57 years, 48.7%
Total urban world population	48 years, 41.0%

How to deal with the missing values?

- We chose to predict missing values because the variables do have some general trend, as we will see in a later slide. We are excluding data sets with over 50% of missing values, which are extreme temperature, mass movement and impact.

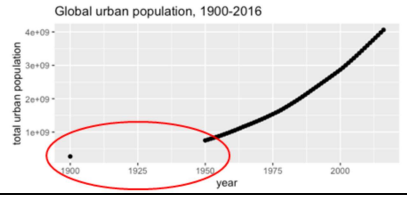
library(imputeTS)

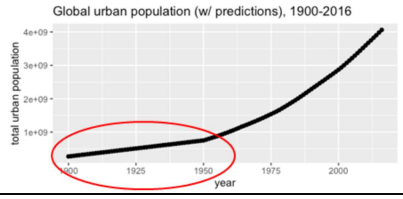
Predict data for time series with the default method: **linear interpolation**

use the **ceiling function to round up** because we believe that the number of recorded natural disasters from our data set often underestimated the true value.

predicted values for urban population:

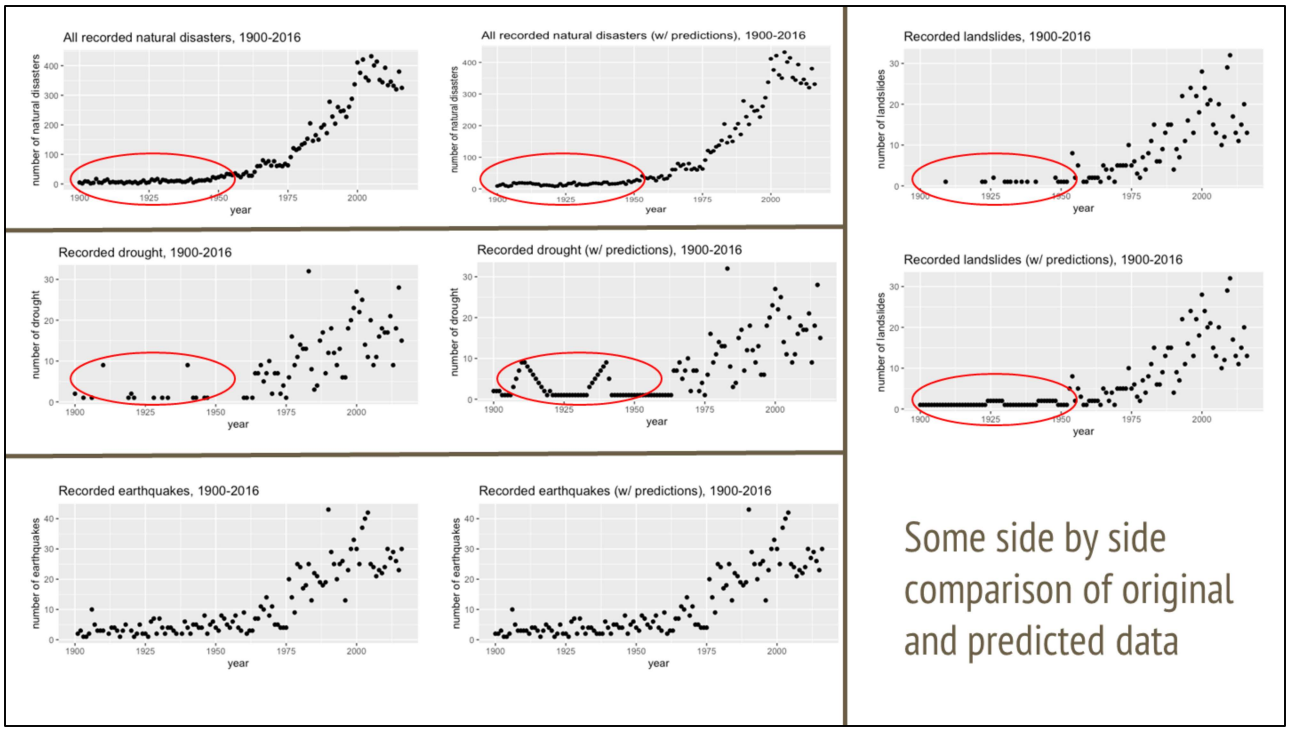
	original	predicted
year <int>	**all.num** <int>	**all.num** <int>
1969	62	62
1970	77	80
1971	61	61
1972	63	65
1973	60	60
1974	68	70
1975	63	63



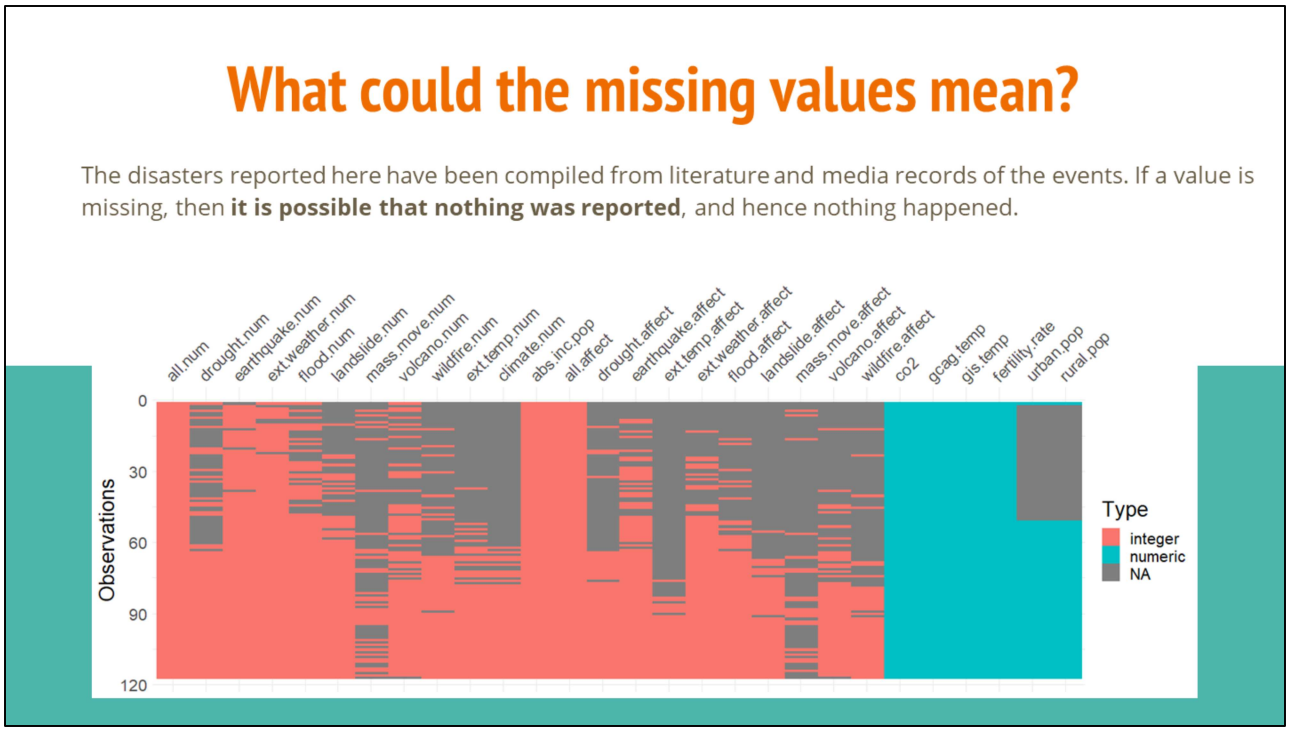


We chose to predict missing values because the variables do have some general trend, as we will see in a later slide. We are excluding data sets with over 50% of missing values, which are extreme temperature, mass movement and impact.

We rounded up the predicted values from linear interpolation because we believe the recorded numbers are likely to be underestimating the true number of natural disasters for the earlier years, and the number must be an integer value.



We have some side by side comparison of the original and predicted data. The predicted data seems reasonable to be used in our analysis.



What could the missing values mean?

- It is possible that the missing values are zeros. The original data sets are compiled from literature and media records of the events. If a value is missing, then it is possible that nothing was reported, and hence a value of zero. We cross referenced with a data set on the number of people affected by the same 10 natural disasters to verify this claim.

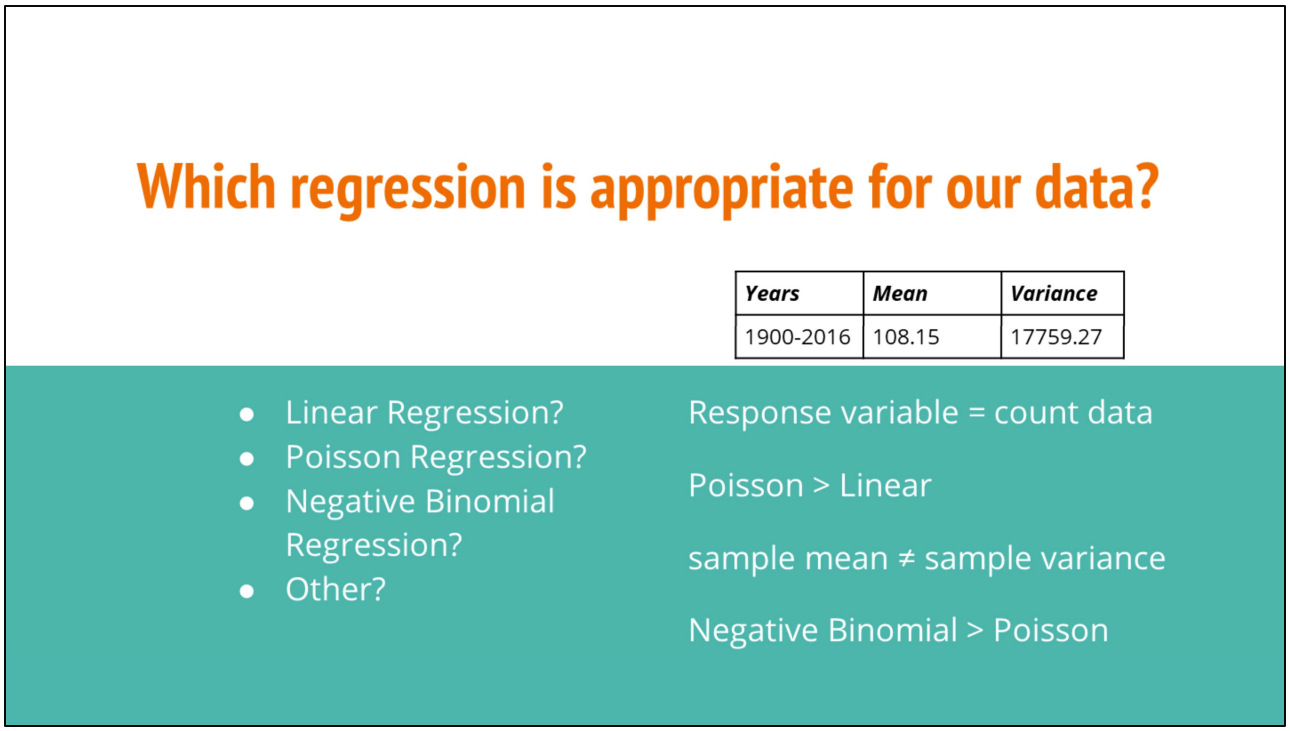
Number of Disasters and Number of People Affected

	Number.of.Droughts	People.affected
1	2	NA
2	NA	NA
3	NA	NA
4	1	NA
5	NA	NA
6	NA	NA
7	1	NA
8	NA	NA
9	NA	NA
10	NA	NA
11	9	32000
12	NA	NA
13	NA	NA
14	NA	NA
15	NA	NA

- There is **no** incident in which no event occured and some people got affected.
- There are some years where a few **disasters happened but no one was affected**. We can see that in the years that no one was affected, very few disasters happened.

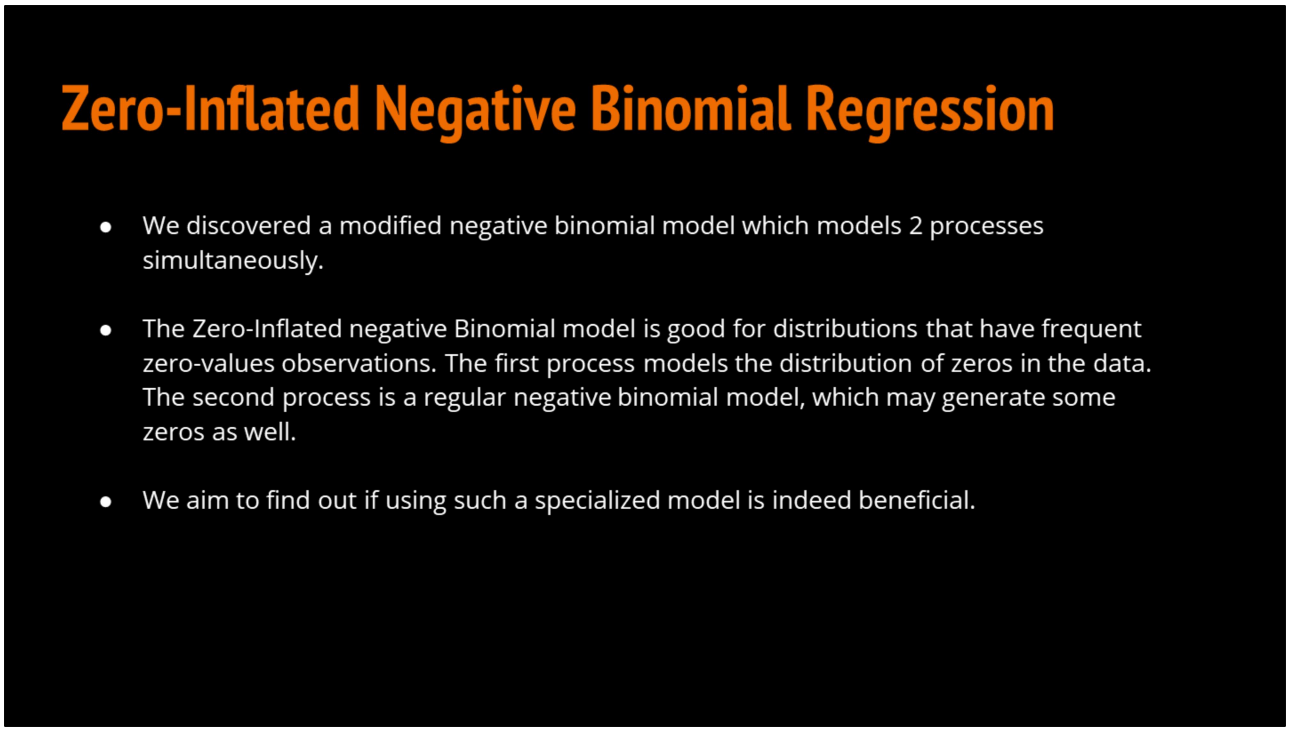
As we see, the missing values in the affected data sets and the missing values in the number variables seem to line up. We can take a closer look at values to verify this.

It can be seen that there is no incident in which no event occured and some people got affected. We note that there are some years where a few disasters happened, but no one was affected, which is reasonable. We can see that in the years that no one was affected, very few disasters happened. Thus, using the missing values as 0 is a possible assumption to use.



Which regression is appropriate for our data?

- Our response variable is count data, so it is common to use a Poisson regression. However, the Poisson regression assumes the mean and variance are the same, which is not true for our data. Thus, we use a negative binomial regression instead.



We found a zero-Inflated negative Binomial model, which is good for distributions that have frequent zero-values observations. The Zero-Inflated negative Binomial model is good for distributions that have frequent zero-values observations. It simultaneously models the distribution of zeros in the data and generates zeros. We tested whether using this specialized model is useful.

To make the predictor ranges more manageable, we scaled the values. We looked at the summary statistics of the zero inflated model, using the “zeroinfl” method from the “pscl” package in R.

Zero-Inflated Negative Binomial Model

Climate disasters include only extreme temperature, extreme weather, flood

log(Climate Disasters) = $\beta_0 + \beta_1$ * (scaled CO2) + β_2* (scaled increase in global population) + β_3 * (scaled absolute urban population)

```
Count model coefficients (negbin with log link):
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)        1.06279    0.11363   9.353  < 2e-16 ***
co2.scaled        -0.55902    0.20122  -2.778  0.00547 **
abs.inc.pop.scaled 0.29379    0.02259  13.003  < 2e-16 ***
urban.pop.scaled   1.14554    0.16355   7.004 2.48e-12 ***
Log(theta)         2.99481    0.20544  14.578  < 2e-16 ***

Zero-inflation model coefficients (binomial with logit link):
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)         3.8880     3.1703   1.226    0.220
co2.scaled        -19.5819    14.6487  -1.337    0.181
abs.inc.pop.scaled  -0.3099     2.0176  -0.154    0.878
urban.pop.scaled   -61.5172 17205.9876  -0.004    0.997
```

- Estimated coefficients of the zero modelling process are **not significant**.
- High p-values suggests that the zero-inflated negative binomial **may not be appropriate**.

The coefficients of the zero-inflated process are far from significant. This suggests that, despite that there are frequent zero-values observations in the data, a separate process may not be required to model their distributions. The coefficients are most likely to be modelling random noise in the data, resulting in high p-values.

Since the coefficients for the negative binomial process are not significant, we plan to use the negative binomial model.

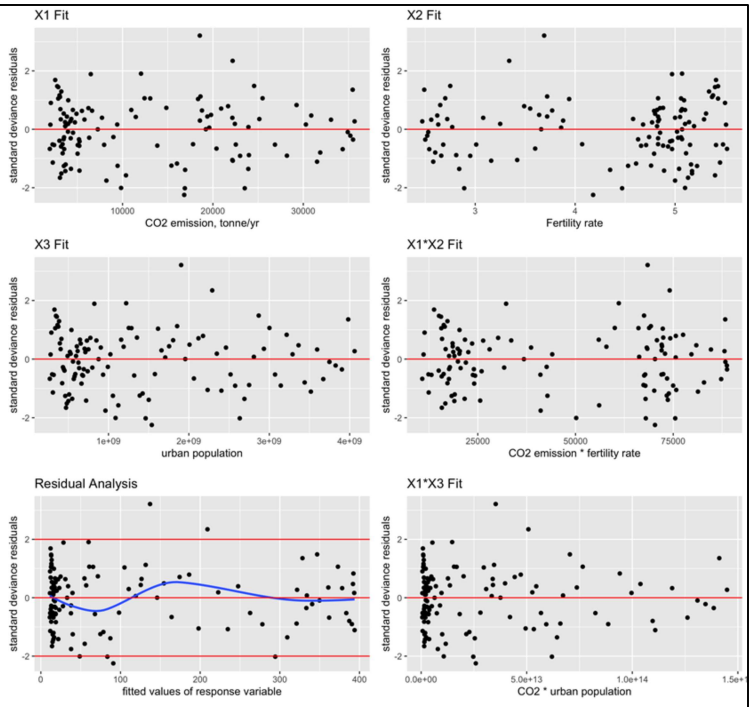
Negative Binomial

Let's turn to the negative binomial regression with the **predicted data**

The **fitted model** we will use:

$$\ln(Yi) = -1.862 + 0.0003129\,(X1i) + 0.6575 * (X2i) + 1.994*10^{-9}\,(X3i) + -0.00004288* (X1iX2i) + -6.536*10^{-14}\,(X1iX3i)$$

X1	CO_2
X2	fertility rate
X3	urban population
X1X2	CO_2 * fertility rate
X1X3	CO_2 * urban population
Y	number of all recorded natural disasters



We are using only CO2, fertility rate, and urban population as our covariates in the negative binomial regression. With interaction terms from CO2 with fertility rate, and CO2 with urban population, the model has a better looking residual plot than without. The fit of our terms with residuals are fairly random and patternless. So we conclude that this model is more appropriate.

Negative Binomial cont.

Holding **all other independent variables constant**:

+1 tonne CO2 = 0.0003129 - 0.00004288 (X2i) - $6.536 * 10^{-14}$ (X3i) in E(ln(Y))

+1 child that would be born to a woman = 0.6575 - 0.00004288 (X1i) in E(ln(Y))

+1 in the global urban population = $1.994 * 10^{-9}$ - $6.536 * 10^{-14}$ (X1i) in E(ln(Y))

Variables		*p-value*	*Significant under =0.05?*
X1	CO_2	$4.23*10^{-7}$	Yes
X2	fertility rate	0.000134	Yes
X3	urban population	$< 2*10^{-16}$	Yes
X1X2	CO_2 * fertility rate	$3.61*10^{-5}$	Yes
X1X3	CO_2 * urban population	$< 2*10^{-16}$	Yes

Bootstrap of 10,000 samples:

Predictors	*Bootstrap 95% Confidence Interval*
CO_2	(0.00021, 0.00043)
fertility rate	(0.336, 0.996)
urban population	$(1.61*10^{-9}, 2.38*10^{-9})$
CO_2 * fertility rate	$(-6.29*10^{-5}, -2.44*10^{-5})$
CO_2 * urban population	$(-7.90e*10^{14}, -5.35*10^{-14})$

- All the terms are significant under 0.05. A change in CO2 depends on the constant values of fertility rate and urban population. And change in fertility rate or urban population depends only on the constant CO2 value.
- A 10,000 bootstrap samples verified the statistical significance of our estimated coefficients and have narrow confidence intervals.

Negative Binomial cont.

Predict for 2020 using outside sources:

[1] IEA (International Energy Agency) reported that CO_2 emission for 2020 can be as low as that of 2010

[2] Directly from the ourworldindata.org dataset with the fertility rate.

[3] In a study by Erin Duffin on Oct 6, 2020, it is estimated that the world urban population would be at 56% for 2020, and the world population now is about 7.8 billion.

364.1007 = **364**
(3 more than that of 2019 from our data set)

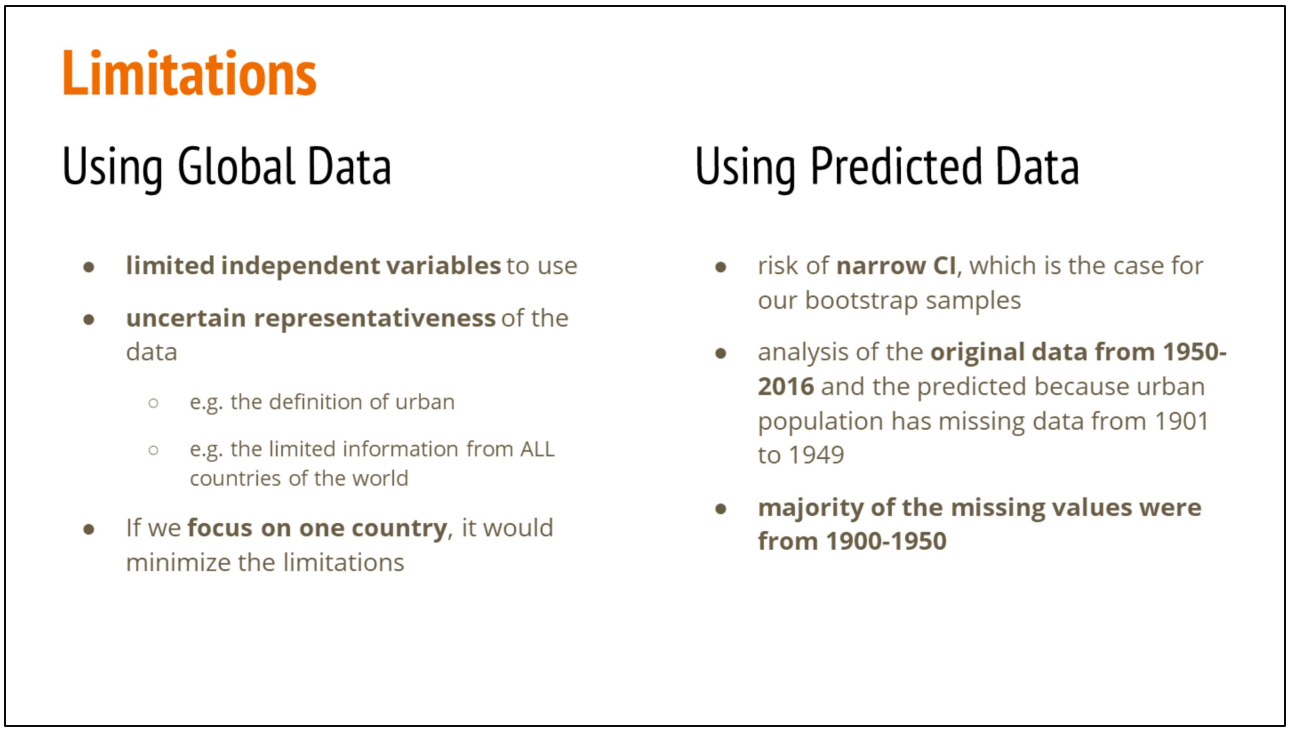
Predictors	*Values*
CO_2 [1]	33066.651
fertility rate[2]	2.44
urban population[3]	4368000000
CO_2 * fertility rate	33066.651 * 2.44 = 80682.63
CO_2 * urban population	33066.651 * 4368000000 = $1.444351*10^{14}$

Using information from the International Energy Agency, the original data sets, and a study of urbanization by Erin Duffin, the predicted number of recorded natural disasters for 2020 is 364, which is 3 more than that recorded for 2019.

Part 3

Implications and Limitations

- Using global data
- Using predicted data
- Using Recorded Data



Using global data

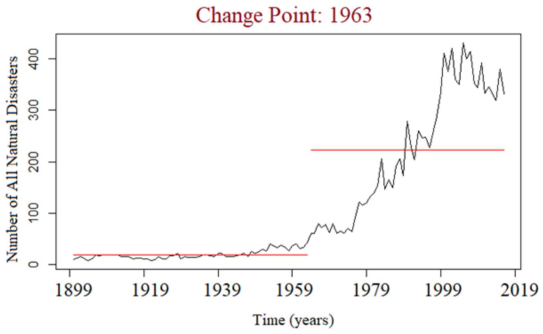
- In using global data, there are limitations on which covariates are available to use. There is also uncertainty in how the global data represents the whole. For example, not all countries may be reporting data with the same definitions. Or there could be under-reporting. In contrast, if we were to focus on a single country, the data would better represent that population.

Using predicted data

- When we use our dataset with predictions for

the missing data, there is the risk of a narrow Confidence Interval. This is true for our bootstrap samples. With more time we would conduct a full analysis of the original and predicted data from 1950-2016 only because the urban population has missing data from 1901 to 1949. However, most of the missing values in the natural disasters data come from 1900-1950, so the results from the two analyses may be similar.

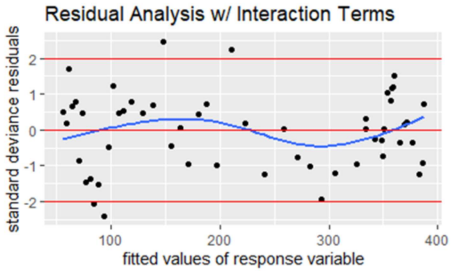
Change Point Analysis



- Used `cpt.meanvar` from `changepoint`
- Change point in the *mean* and *variance* with an uncertainty of ±1 year.
- Predicts 653 disasters for 2020.

	Variable	*z*	*p-value (<0.10?)*
X_1	CO_2	1.867	0.0619
X_3	Urban.pop	14.283	$< 2*10^{-16}$
X_1X_3	CO_2 * Urban.pop	-10.603	$< 2*10^{-16}$
Y	Number of all recorded natural disasters		

$$Y = 1.719 + X_1(4.056 \cdot 10^{-5}) + X_3(2.027 \cdot 10^{-9}) - X_1X_3\ (3.762 \cdot 10^{-14})$$



To try to improve the fit, we used Change Point Analysis which identifies when the distribution of the data changes. We perform this analysis because the availability of information has changed over time. For this study, we wanted the change point to be determined by there being both a change in the mean and a change in the variance. To do this, we used the function `cpt.meanvar` from the library `changepoint` with a 95% confidence level.

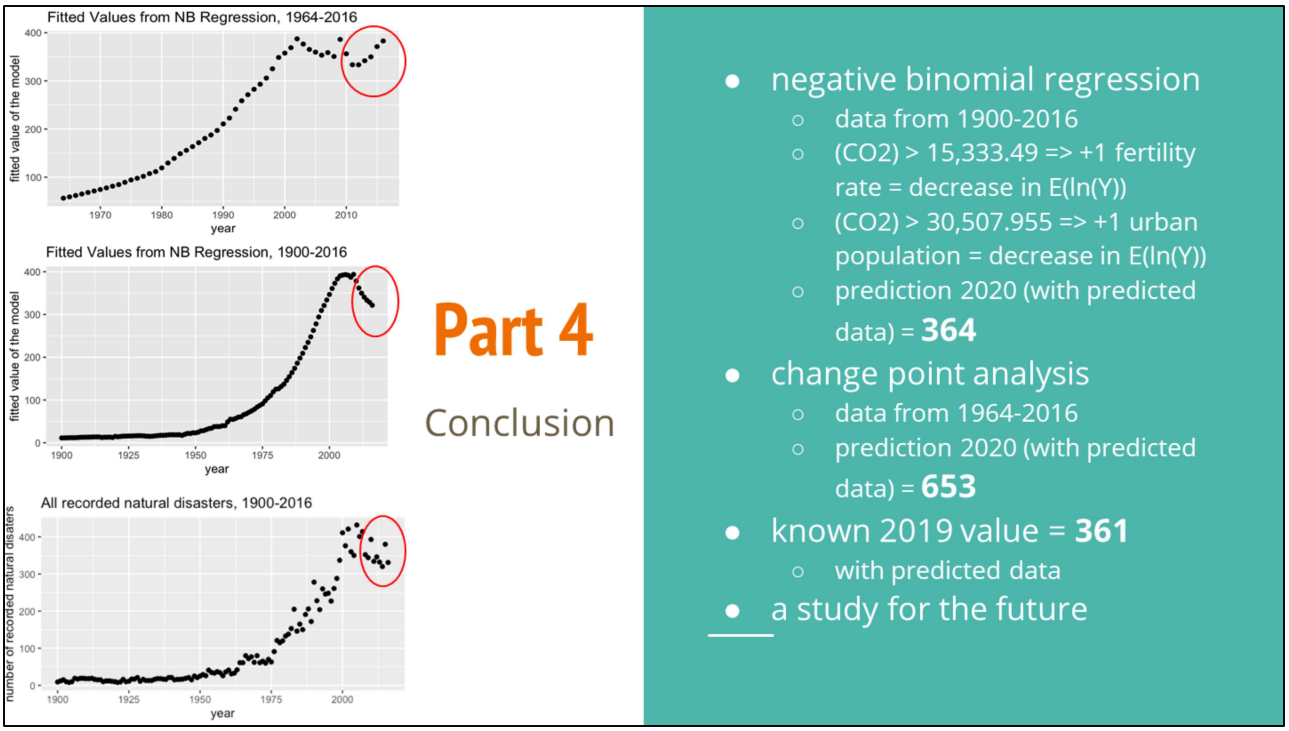
For the raw data, the result of the analysis produced a change point of the year 1962. Performing it on the data with predictions, we get a

change point of the year 1963, which leaves us an uncertainty of at least a year for when the actual change point occurred.

We can consider the predicted data from 1963-2016, which has 53 data points, for our fit. It could give us a more accurate prediction of how many natural disasters there are in the present and how that is changing with respect to the covariates. The negative binomial fit was first used with the original model of 3 covariates with 2 interaction terms. The p-value for the X2 parameter of fertility rate was 0.29 which is greater than the significance level of 0.10, thus we conclude the null hypothesis that the coefficient is zero. Meaning, we can remove X2 and its interaction term from the fit. The updated fit is the one displayed in the table on this slide. We see that the updated model is acceptable because their p-values are all under the 10% significance level.

A standardized residual fit was performed for our model. For a 90% confidence level, the residuals should fall within just a little more than 2 standard deviations from the mean. This is visible in our plot, and we see that the points appear randomly scattered about zero such that the smoothed fit hugs zero.

Using the fit with the estimated CO2 concentrations and the estimated global urban population, 653 natural disasters are predicted for 2020.



Let's move onto our results. Holding all other independent variables constant in our model, if the value for CO2 is large enough, the other two covariates yield a decrease in the expected log of the number of natural disasters. We can also see the decreasing trend since 2000.

This trend is consistent with the fitted values of the model from 1900-2016 and our prediction of 364 disasters is reasonable under this trend. However, per our model fitted from 1964 to 2016, the prediction for 2020 is 653, about 1.7 times that of 364. We also see that the predicted model from 1964 is increasing since 2010, which does not

match the trend in the original data.

In conclusion, our analysis from 1900 to 2016 is more reasonable than that from 1964 to 2016. Until we have more data to work with, analysis on data since 1900 is promising. In accordance with our change point analysis, our covariates can be valid for data from 1964 into the far future.

Work Cited

Absolute increase in population. Our World in Data. Web. 23 Oct 2020. https://ourworldindata.org/grapher/absolute-increase-global-population?time=1900..latest.

3. *Children per woman.* Our World in Data. Web. 23 Oct 2020. https://ourworldindata.org/world-population-growth.

CO2 emissions by region. Our World in Data. Web. 23 Oct 2020. https://ourworldindata.org/co2-emissions#year-on-year-change-in-global-co2-emissions.

2. Duffin, Erin. "Urbanization by Continent 2020." *Statista*, 6 Oct. 2020, www.statista.com/statistics/270860/urbanization-by-continent/.

Global reported natural disasters by type. Our World in Data. Web. 23 Oct 2020. https://ourworldindata.org/natural-disasters.

1. Iea. *Global Energy and CO2 Emissions in 2020 – Global Energy Review 2020 – Analysis*. Apr. 2020, www.iea.org/reports/global-energy-review-2020/global-energy-and-co2-emissions-in-2020.

Number of people affected by natural disasters. Our World in Data. Web. 23 Oct 2020. https://ourworldindata.org/natural-disasters.

Urban and rural population projected to 2050. Our World in Data. Web. 23 Oct 2020. https://ourworldindata.org/urbanization.

Van Aalst, Maarten K. "The impacts of climate change on the risk of natural disasters. Disasters", vol. 30, 2006, p. 5-18. doi:10.1111/j.1467-9523.2006.00303.x

"Global Greenhouse Gas Emissions Data." *Greenhouse Gas Emissions*, EPA, 10 Sept. 2020, www.epa.gov/ghgemissions/global-greenhouse-gas-emissions-data.