Trends of Crime in the USA

Introduction

Crime is something that is unpredictable and once it happens, it is too late. Predicting the exact time or type of crime committed is impossible, but what if we can determine if crime is more likely to happen under certain conditions? I would like to determine if the number of crimes is related to other factors such as population, weather, gas prices, health, income, or location. My hypothetical client would be any kind of law enforcement agency or government in general. They would care about this particular issue because they can assign more officers on patrol or be warier when certain observations are met. For an example, if it was discovered that after a major disaster such as a hurricane struck the number of theft-related crimes increased in that area. Increasing the number of officers on patrol in that particular area could help lower that number. Or if it was found that the average number of days of poor mental health in a month is correlated with a high number of crimes, allocating funds to mental health programs could indirectly lower crime rate.

To help with exploratory analysis, I will create an interactive visualization using the Dash package from Python. This interactive visualization could be used by a law enforcement agency to explore further or by a person who wishes to do additional research in this area. Next, I will build a statistical model to determine what factors best contributes to the total number of crimes.

Datasets

Crimes

I will be using multiple datasets from a variety of different sources.

The first type of data I will use is crime data. One dataset is located at https://crime-data-explorer.fr.cloud.gov/downloads-and-docs
It is the Summary(SRS) Data with estimates CSV file under the additional datasets. It contains aggregated crime count data at the state and national level from 1995 to 2016.

It includes variables such as year, state, population, violent crimes, homicide, rape, robbery, aggregated assault, property crime, burglary, larceny, and motor vehicle theft.

Weather

The second type of data I will use is weather data. I manually collected 50 datasets, one from each state in the USA at https://www7.ncdc.noaa.gov/CDO/CDODivisionalSelect.jsp#
I selected the “State” tab, with the period from 01/1994 to 12/2017, and comma delimited text output. I saved each file as the name of the state.

The variable names can be found at the bottom of the page under Indices Output. A more extensive variable definition is available at: ftp://ftp.ncdc.noaa.gov/pub/data/cirs/climdiv/divisional-readme.txt
I am mainly interested in Year/Month, TMIN, TMAX, and TAVG. The latter three are in Fahrenheit.

Storm Events

I use another dataset located at ftp://ftp.ncdc.noaa.gov/pub/data/swdi/stormevents/csvfiles
I downloaded the StormEvents_details file from 1994 to 2017. Some of the variable definitions can be found at the top in the Storm Data Export Format word document.

The variables of interest here other than the Year and Date are: EVENT_TYPE, INJURIES_DIRECT, INJURIES_INDIRECT, DEATHS_DIRECT, DEATHS_INDIRECT, DAMAGE_PROPERTY, and DAMAGE_CROPS. EVENT_TYPE is the type of event such as 'Tornado'. The other variables are self-explanatory.

Standard Precipitation Index

I use another dataset located at https://ephtracking.cdc.gov/download
I downloaded the Standard Precipitation Index (Pearson) file. It includes county-level data from 1895 to the present.

The only unique variable of interest is the Standard Precipitation Index (SPI).

Population

The third type of data I will use is population data. The dataset is located at https://seer.cancer.gov/popdata/download.html#19
I downloaded the adjusted 1969-2016 White, Black, Other data County-level data. The adjusted data takes into account Hurricane Katrina and Rita.

The variable dictionary can be found at: https://seer.cancer.gov/popdata/popdic.html
The unique variables of interest are: population, race, sex, and age.

Income

The fourth type of data I will use is Income data. The dataset is located at https://www.census.gov/data/tables/time-series/demo/income-poverty/historical-income-households.html
I downloaded the Table H-8 data representing the median household income by state. The data ranges from 1984 to 2016.

The only unique variable here is the median income of households.

Gas Prices

The fifth type of data I will use is Gas price data. The dataset is located at
https://www.eia.gov/state/seds/seds-data-complete.php?sid=US
Under data files, Prices and expenditures, 1970-2015, all states, I downloaded the CSV file for Prices.

The unique variable of interest is the gas price in dollars per million british thermal units.

Health

The last type of data I will use is the Behavioral Risk Factor Surveillance System data. It is available at
https://www.cdc.gov/brfss/annual_data/annual_1994.htm
For the year 1994. I use the data ranging from 1994 to 2016. I downloaded the XPT transport SAS files under each year. The extensive variable definitions are also viewable from this page.

Data Cleaning

Data cleaning is divided into two steps. Full details are in the jupyter notebooks.

Part 1

Weather

I load in each csv file using pandas, concatenating each consecutive data frame onto the previous. I Created a new column to store which what state the data represents.

I subsetted the data to include only relevant columns: State, YearMonth, TVG, TMIN, and TMAX.

I extracted the year and month fro the YearMonth column as separate columns and stored YearMonth as a datetime variable. I converted each column into their proper datatype and saved the cleaned data into a new csv file called Weather.

Income

I loaded in the data and noticed that the years were represented as individual columns. I melted the data frame and stored the years all into one column. I fixed the datatypes of each column and converted the date column to a datetime object. The cleaned data frame was saved to a new csv file called Income.

Population

I loaded in the data and used the variable dictionary to determine how to extract the data. The data was stored into one column and had no delimiter. I created new columns for each relevant variable: Year, State_Abbrev, Race, Sex, Age, and Population.

I subsetted the data to include only years past 1993. I created a new column called “Age_Group” to pool the given age categories into fewer categories. Ages from 0-19 are “Child/Teen”, 20-39 are “Young Adult”, 40-59 are “Older Adult”, and the last one is 60+. I corrected all datatypes.

First, I grouped by Year, State_Abbrev, Race, Sex, and Age_Group finding the aggregate sum Population. This will be used later and grouped in different ways to create different types of visualizations. I converted the multi indexed data frame into a regular data frame with each variable in its own column. I saved this to a csv file called Population2.

Next, I grouped by Year and State_Abbrev finding the aggregate sum Population. This is the data frame that would be merged to the other datasets . I saved this to a csv file called Population

Storm Events

I loaded in the data files using a for loop, concatenating each consecutive data frame to the previous.

The data was subsetted to include only relevant variables: BEGIN_YEARMONTH, STATE, YEAR, EVENT_TYPE, INJURIIES_DIRECT, INJURIES_INDIRECT, DEATHS_DIRECT, DEATHS_INDIRECT, DAMAGE_PROPERTY, and DAMAGE_CROPS.

I extracted the month from the BEGIN_YEARMONTH column and stored it in a column called Month. All null values were dropped from the data frame so that the DAMAGE_PROPERTY and DAMAGE_CROPS columns could be cleaned.

Both of these problem columns contained values such as “10M”, “3.00K”, “K”, “05”, etc. I checked through the value_counts to find all the cases to account for. In general, I removed the letter from each row and depending on what letter was removed, the number remaining was multiplied by a number. For “K”, the multipler was 1000 for an example. If there was only a “h” in the column, it was treated as 0 because there were values that were “1.00K” meaning that the “h” did not represent 100 otherwise it would be written as “1.00h”.

Incorrect datatypes were converted to their proper types and the cleaned data frame was saved to a csv file called Storm_Events

Gas Prices

I loaded in the data and subsetted the data so that it included only ‘MGACD’ under the MSN column. This represents motor gas usage. Since the years were stored in multiple columns, I melted it so that it

was stored in one column. The data was subsetted again to include only years past 1993. The numbers in the Gas column represented $ per million btu (british thermal unit). I converted it into $ per gallon.

I deleted irrelevant columns, fixed column labels to match other datasets, and saved the cleaned data frame to a csv file called Gas_Price.

Crimes

I loaded in the the data and included only non null values of state_abbr. This removed the nationwide aggregated data. I fixed column labels so that they match other datasets of the same information.

I compared the population value for 1995 in this data with the Population data. They don't match. I deleted this population column and will use the population values in the other data frame since the population came from a more reliable source. I removed the rape_revised column because there is already another column containing counts of rape. The cleaned data was saved to a csv file called Crimes.

Standard Precipitation Index

I loaded in the data and subsetted the data to include only years past 1993. I removed the 'fips' column before melting the data. Months were stored in multiple columns instead of one. The month column included months in their three letter abbreviation. I made a function to change them into their respective number representations.

I grouped by state, year and month and aggregated using the mean for SPI values. I rename columns to match other datasets and then save the cleaned data frame to a csv file called SPI.

Health

I loaded in the XPT files using pandas read_sas. The data was immediately subsetted to include only the variables: _STATE, PHYSHLTH, MENTHLTH, and Year. I removed certain values of _STATE so that values such as "virgin islands" was not included as a state. I created three lists to help load in the data and clean it. Years was created to match the data from 1994 through 2016. 'states' was created with the states in alphabetical order to help create a State column to extract the correct states for each observation. The 'numbers' column was created to help determine what state each observation was. The _STATE variable contained numerical values which represented a certain state. The numbers used skipped numbers frequently. The numbers in the numbers list I created are listed with their state representation in alphabetical order.

The data was further subsetted so that it includes MENTHLTH and PHYSHLTH values 30 and below. This excludes responses such as "refuse to answer" and "don't know". I aggregated by mean and grouped by Year and State. Each consecutive XPT file was stripped down to the point of the aggregated data frame before merged into one. The resulting data frame is saved to a csv file called Health.

Part 2

I loaded in Weather, Income, Population, Population2, Storm_Events, Gas_Prices, Crimes, SPI, and Health.

Weather

I created two separate data frames. One was the mean aggregate grouped by Year and State. The other one was the Month data. They are saved to data frames called Weather_Year and Weather_Month.

Population

I renamed the values in the columns to indicate what they represent. I grouped them by Year, State, Race, Sex, Age, and Age Group to sum up the populations for those categories. I saved the resulting data frame to a file called 'Population_Groups.csv'.

Storm Events

I renamed the columns to match the other datasets. I converted the EVENT_TYPE column to several dummy variables. I combined this result with the original data frame and deleted the EVENT_TYPE column. I made two separate data frames as I did with weather. One contains the year level data and the other one contains the month aggregated data by state.

SPI

I renamed the state column to match other datasets. I saved the original dataset as the month aggregated data and made the year aggregated data as a separate data frame.

Year Combined

I used a left join to combine on 'State' and 'Year' for Weather_Year, Storm_Year, and Health.

I created a dictionary of the states with their abbreviation as the values. I converted this into a data frame and merged the previous data frame with this one on 'State' using a left join so that only the states in the original data frame are included.

I merged the remaining datasets on 'State_Abbrev' and 'Year'. I saved the resulting data to a file called Year_df.

Month Combined

I repeated similar steps with the Month aggregated data. This only included Weather_Month, Storm_Month, and SPI_Month. The resulting csv file is called Month_df

Visualization Application Description

Crime Type Checkbox

What crimes to include in the scatter plots and overall time series. Depending on what is selected, you can find relationships between all crimes, a subset of crime types, or a single crime type.

Select Variables

There is a linear or log option and a average or sum option. The dropdown menu selects what variable to plot a time series and a scatter plot of that variable versus the total number of crimes committed from selected crimes in the checkbox. The sum option will compute the sum of that selected variable from every selected state. The average option will compute the average of that selected variable depending on what states are selected. The linear/log option changes the scale of the scatter plot so that it is easier visualized when points are clustered.

Choropleth map

Hovering over each state will list the total number of crimes committed depending on what is selected for crime type. It shows how many of each crime contributes to the overall number of crimes for that state. It is heat mapped as well to easily visualize which states have a higher number of crimes committed.

Variable Time Series

Shows the distribution of selected variables over time averaging or summing each state at each year depending on selected states. There are two time series so that you can compare one variable to another with ease.

Scatter Plots

Shows the scatter plot of the variable versus the total number of crimes committed based on selected crime types. You can use this to visualize each variable against violent crime, rape, etc individually. You can also use this to visualize each variable against a subset of crimes or all of them. Two scatter plots are included to to able to visualize two variables at once against total number of crimes committed. You can hover over the points to determine what year and state that point is.

Select States

This is a multi-select dropdown menu. Depending on what is selected affects the appearance and results of all visualizations and the table at the bottom.

Overall Time Series

Plots the total number of crimes from selected crimes over time. There is a slider to select the range of desired years. There is also an 'Input Year to draw line' option to draw a vertical line. This may be helpful in case there are certain years with important events to see before/after effects of the event. For an example, Hurricane Katrina was in 2005. We can draw a vertical line at the year 2005 to indicate this event and easily visualize trends before 2005 and after 2005.

Table

We can filter additional variables and change what data is included in the time series and scatter plots. For an example, filtering median income > 40000 and then visualizing the average number of days in a month of poor mental health versus crimes committed or just the time series of mental health over time.

Example Usage

https://github.com/Skywind555/Springboard/blob/master/Project%202/Example%20Use%20App.ipynb

Statistical Model

I chose to model the data with a negative binomial model because I'm trying to model total number of crimes committed which is based off of counts. Theoretically, it is better than the poisson model because we don't have negative counts.

We can model it using a general linear model with a log link function.

$\log \mu = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots$ where μ is the expected count of the response.

Univariate Analysis:

I created our response variable, total_crimes, with all of our crimes summed together. I began with univariate analysis by looking for significant p values trying to model total_crimes with each of the other variables in the dataset.

The following were significant minus overlap: Year, State, TAVG, TMIN, TMAX, Median_Income, INJURIES_DIRECT, INJURIES_INDIRECT, DEATHS_DIRECT, DEATHS_INDIRECT, DAMAGE_PROPERTY, DAMAGE_CROPS, Avalanche, Blizzard, Debris Flow, Dense Fog, Dense Smoke, Drought, Dust Storm,

Extreme Cold/Wind Chill, Flash Flood, Frost/Freeze, Funnel Cloud, Hail, Heavy Snow, High Surf, Hurricane (Typhoon), Lake-Effect Snow, Lakeshore Flood, Lightning, Rip Current, Sneakerwave, Storm Surge/Tide, Strong Wind, Thunderstorm Wind, Tornado, Tropical Storm, Tsunami, Waterspout, Wildfire, Winter Storm, Population, Gas_Per_Gallon, and SPI

Building a Multivariate Model:

I performed forward selection using lower aic as the criteria.

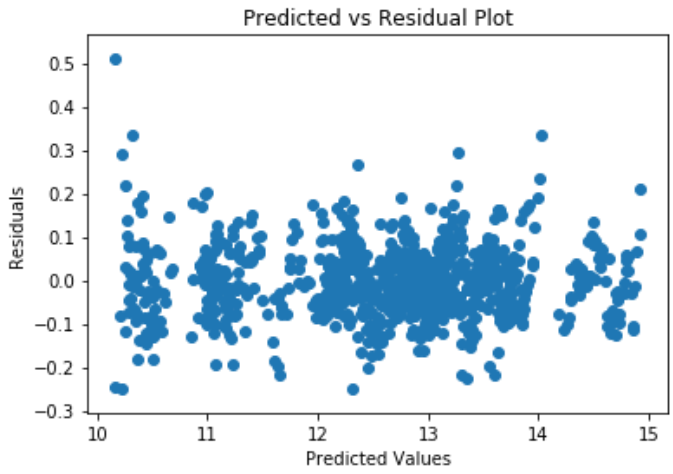
I ended up with a model with State, DEATHS_DIRECT, Gas_Per_Gallon, SPI, Population, and Year as the predictors.

Checking Model Assumptions:

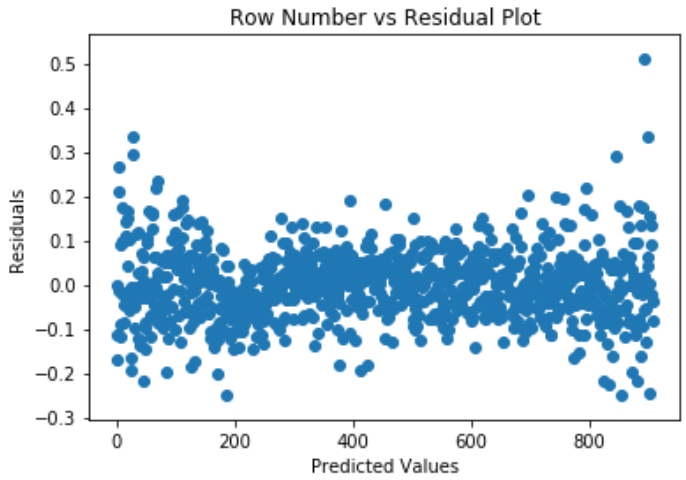
The negative binomial model assumes the conditional means are not equal to the conditional variances. We can test this with a log likelihood ratio test with the Poisson model since the dispersion parameter is held constant in the Poisson model. The Poisson model is nested inside the negative binomial model, so it is the reduced model.

I find that the negative binomial model is strongly favored which shows that the conditional means are equal to the conditional variances.

Since we have a log link function, total_crimes should depend linearly on the predictors.



Checking the residual plot above, there are no signs of non-linearity so our assumptions are satisfied.



Checking the row number versus residual plot, we see that the independence assumption is met as well since most of the points are situated around 0.

Checking For Collinearity:

	Year	Gas_Per_Gallon	DEATHS_DIRECT	SPI	Population
Year	1.000000	0.889255	0.243404	0.010347	0.052535
Gas_Per_Gallon	0.889255	1.000000	0.254494	0.026557	0.029131
DEATHS_DIRECT	0.243404	0.254494	1.000000	-0.023925	0.266411
SPI	0.010347	0.026557	-0.023925	1.000000	-0.134993
Population	0.052535	0.029131	0.266411	-0.134993	1.000000

Gas_Per Gallon is highly correlated with Year, which may not be a problem.

Checking our Variance Inflation Factors, Population has a value of 159.71. I suspect that it is correlated with State.

I compare their deviance values as separate univariate models to determine which predictor is better. A univariate model with State as the predictor results in a deviance of 14.088 while a univariate model with Population as the predictor results in a deviance of 335.05. Since we want to minimize our deviance, I choose to remove Population from our multivariate model.

I also notice that the p value for DEATHS_DIRECT and SPI are above 0.05 and their associated parameter estimates are very low, making them practically insignificant as well. I try to remove them one at a time comparing to see if aic decreases. I find that removing both of them reduces the aic.

Our final model is:

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	54.0225	1.980	27.291	0.000	50.143	57.902
State[T.Arizona]	0.4028	0.026	15.720	0.000	0.353	0.453
State[T.Arkansas]	-0.5033	0.026	-19.657	0.000	-0.554	-0.453
State[T.California]	1.8994	0.026	73.978	0.000	1.849	1.950
State[T.Colorado]	-0.1289	0.026	-5.030	0.000	-0.179	-0.079
State[T.Connecticut]	-0.7041	0.031	-22.675	0.000	-0.765	-0.643
State[T.Delaware]	-1.7206	0.026	-66.338	0.000	-1.771	-1.670
State[T.Florida]	1.4783	0.026	57.735	0.000	1.428	1.529
State[T.Georgia]	0.7263	0.026	28.354	0.000	0.676	0.776
State[T.Idaho]	-1.6107	0.028	-58.380	0.000	-1.665	-1.557
State[T.Illinois]	0.8586	0.026	33.520	0.000	0.808	0.909
State[T.Indiana]	0.1669	0.026	6.437	0.000	0.116	0.218
State[T.Iowa]	-0.7858	0.026	-30.687	0.000	-0.836	-0.736
State[T.Kansas]	-0.5715	0.026	-22.317	0.000	-0.622	-0.521
State[T.Kentucky]	-0.4981	0.026	-19.447	0.000	-0.548	-0.448
State[T.Louisiana]	0.1128	0.027	4.164	0.000	0.060	0.166
State[T.Maine]	-1.7349	0.029	-58.986	0.000	-1.793	-1.677
State[T.Maryland]	0.1716	0.026	6.696	0.000	0.121	0.222
State[T.Massachusetts]	-0.0359	0.030	-1.192	0.233	-0.095	0.023
State[T.Michigan]	0.6149	0.026	24.014	0.000	0.565	0.665
State[T.Minnesota]	-0.1608	0.026	-6.277	0.000	-0.211	-0.111
State[T.Mississippi]	-0.6272	0.028	-22.771	0.000	-0.681	-0.573
State[T.Missouri]	0.2231	0.026	8.712	0.000	0.173	0.273
State[T.Montana]	-1.8230	0.026	-71.104	0.000	-1.873	-1.773
State[T.Nebraska]	-1.1370	0.026	-44.387	0.000	-1.187	-1.087
State[T.Nevada]	-0.6546	0.028	-23.277	0.000	-0.710	-0.599
State[T.New Hampshire]	-1.8258	0.029	-62.164	0.000	-1.883	-1.768
State[T.New Jersey]	0.2165	0.026	8.454	0.000	0.166	0.267

State[T.New Mexico]	-0.7345	0.026	-27.941	0.000	-0.786	-0.683
State[T.New York]	0.9796	0.026	38.227	0.000	0.929	1.030
State[T.North Carolina]	0.6670	0.026	25.719	0.000	0.616	0.718
State[T.North Dakota]	-2.5226	0.026	-95.912	0.000	-2.574	-2.471
State[T.Ohio]	0.7955	0.026	30.261	0.000	0.744	0.847
State[T.Oklahoma]	-0.2282	0.027	-8.562	0.000	-0.280	-0.176
State[T.Oregon]	-0.2352	0.026	-9.049	0.000	-0.286	-0.184
State[T.Pennsylvania]	0.5685	0.026	22.182	0.000	0.518	0.619
State[T.Rhode Island]	-1.8347	0.031	-59.113	0.000	-1.895	-1.774
State[T.South Carolina]	0.0893	0.026	3.488	0.000	0.039	0.140
State[T.South Dakota]	-2.3758	0.026	-91.571	0.000	-2.427	-2.325
State[T.Tennessee]	0.3419	0.026	13.014	0.000	0.290	0.393
State[T.Texas]	1.6798	0.026	65.601	0.000	1.630	1.730
State[T.Utah]	-0.6637	0.026	-25.245	0.000	-0.715	-0.612
State[T.Vermont]	-2.5253	0.029	-87.943	0.000	-2.582	-2.469
State[T.Virginia]	0.0988	0.026	3.761	0.000	0.047	0.150
State[T.Washington]	0.4095	0.026	15.952	0.000	0.359	0.460
State[T.West Virginia]	-1.4076	0.028	-50.047	0.000	-1.463	-1.352
State[T.Wisconsin]	-0.1701	0.026	-6.635	0.000	-0.220	-0.120
State[T.Wyoming]	-2.4785	0.026	-94.331	0.000	-2.530	-2.427
Gas_Per_Gallon	0.0403	0.007	5.732	0.000	0.027	0.054
Year	-0.0206	0.001	-20.692	0.000	-0.023	-0.019

Model Interpretation:

For one unit increase in Year, the difference in the log of expected counts of the total number of crimes committed in the USA decreases by 0.0206. Similar conclusions can be made about the other parameter coefficients.

Conclusion / Limitations

In the end, I only found that the State, Gas_Per_Gallon, and Year were good predictors for total crimes committed while testing at the 0.05 level. Since Gas_Per_Gallon is heavily correlated with the Year, practically our results don't help us learn more about the conditions when crime count increases. We

know from the brief exploratory data analysis that each state has its own distribution of total crimes committed which is correlated with Year and Population as well.

Our final model is not optimized for prediction as we can see with the very high intercept coefficient. There may be a better model given the data, but only forward selection with lower aic as the critera was explored. Interactions were not explored either and may have been good to explore considering the complex relationship between state, population, year, gas_per_gallon, and total_crimes. It also may have been good to try to add some of the non significant predictors at the univariate level as they may need the effects of other variables to appear significant. Variables such as MENTHLTH had a low p value for the univariate model, but looking at our exploratory data analysis, it definitely seemed to have an impact on total crimes committed.

Since we have state and year aggregated data, we are not showing that having a higher gas price for one day increases the probability of crime occurring. We are showing that years with higher gas prices on average increases the total number of crimes committed. Despite knowing that, it is unclear is more crimes were committed for parts of the year where gas prices were higher on average or lower on average. There could also be lurking variables here.

Overall, it seems more research needs to be done in order to determine what factors increase the total number of crimes other than what is obvious considering the very complex relationships between our variables.

Future Work

In a future study or rework of the current analysis, it may be worth trying other methods of building the model. It may also be good to look at responses other than total crimes committed. Instead, a multivariate model for each of the crime types instead of an overall crimes committed would be ideal.

However, a multivariate time series model would be most ideal and most likely result in a better model since it is time data.

Another option is to try modeling a proportion of crimes committed relative to population size instead of crime count.

More data can also be collected to be merged with the current data, but may need certain government clearances or other types of clearances to obtain the data.