The effects of co-morbidities on COVID-19 cases and deaths

Based on U.S. County Data

Vilius Smilinskas

5/30/2020

Statistical Question

- COVID-19 is a respiratory virus that is well known to affect different sections of the population with different severity.
- Based on a study of 5700 patients “The most common comorbidities were hypertension (3026, 56.6%), obesity (1737, 41.7%), and diabetes (1808, 33.8%)” (https://jamanetwork.com/journals/jama/fullarticle/2765184)
- By utilizing data on county total number of cases and deaths on a certain day, as well as health rankings for counties around the country we were able to examine some of the major contributors.

Variables

County Cases and Deaths (covid-19-county-level-data.csv)

- date : The date of the record.
- county: The county of the record.
- state: The state in which the county is in.
- cases: Number of total cases.
- deaths: Number of total deaths.
- Since the data provided is the total number on that date, we will also be calculating and examining two new variables (casenew, deathnew)

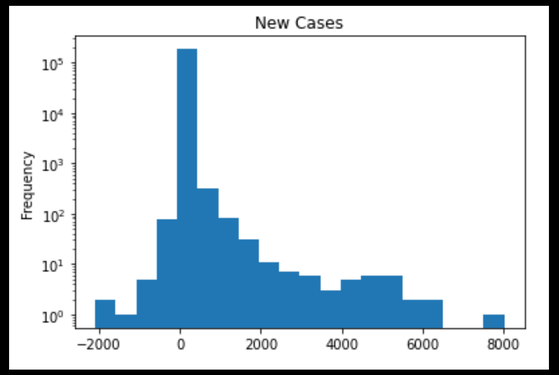
Variables

County Health Rankings (us-county-health-rankings-2020.csv)

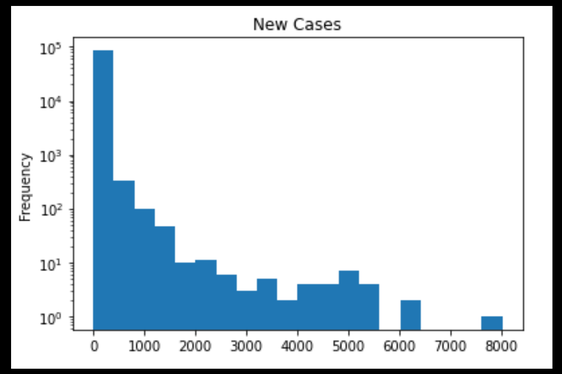
- By joining our data we will compade the county health data to their overall number of cases and deaths. The health variables will function as the predictive variables for our future models.
- county: the county of the record.
- state: the state in which the county is in.
- percent_fair_or_poor_health : Percentage of the population that is considered in fair or poor health.
- percent_smokers: Percentage of the population that smokes.
- percent_adults_with_obesity: Percentage of the population that is obese.
- percent_excessive_drinking: Percentage of the population that are excessive drinkers.
- income_ratio: The income ratio among the state population
- percent_adults_with_diabetes: Percentage of the population that have diabetes.

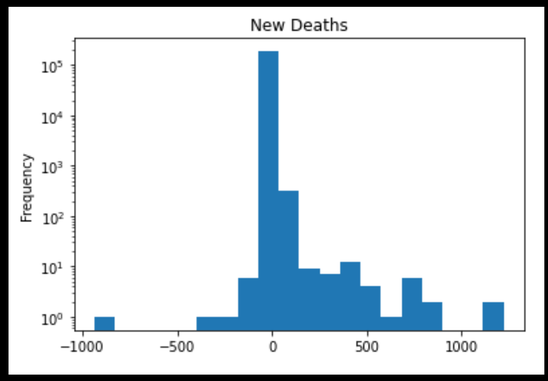
Variable Histograms

Both data sets

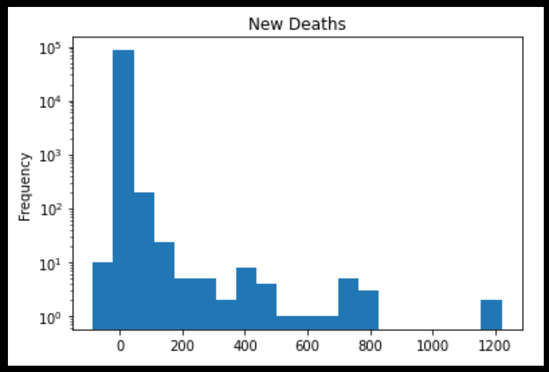


When exploring the new cases histogram we see that there are vales below zero. Exploring the data set we can find that some total county values of cases are adjusted and reported at a lower value thus we are going to be excluding the negative values. Because there is no way to decrease the number of confirmed cases.

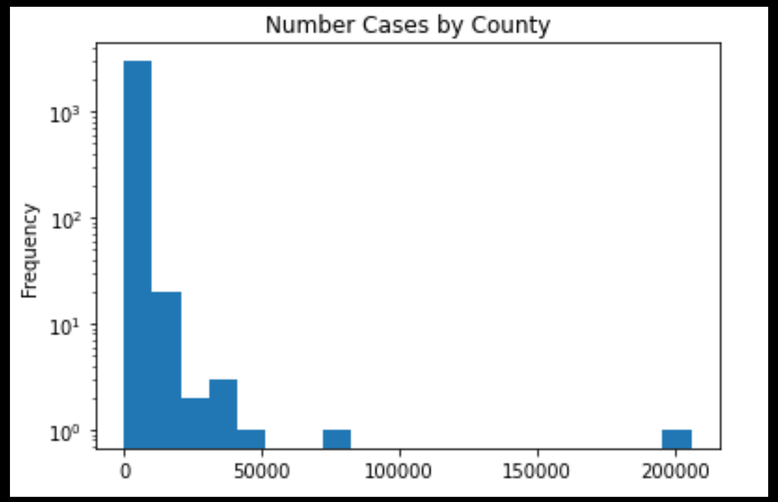




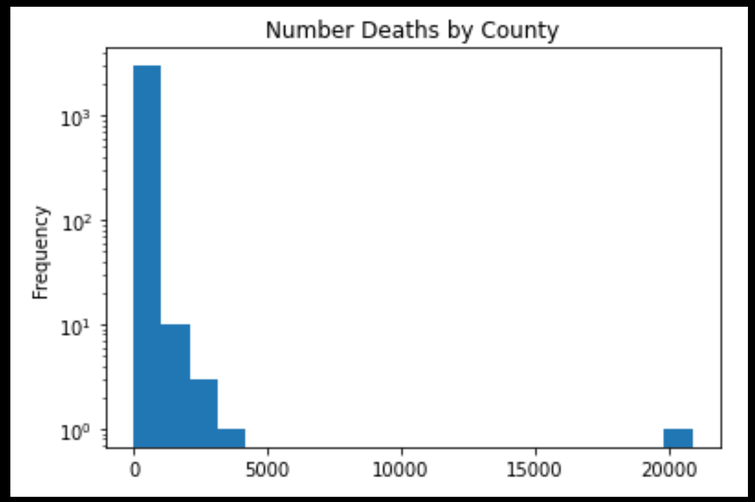
Similarly due to the same situation discussed, there are negative values in our data set. But the excluded rows of data from the condition above fixes the graph already.



Variable Histograms

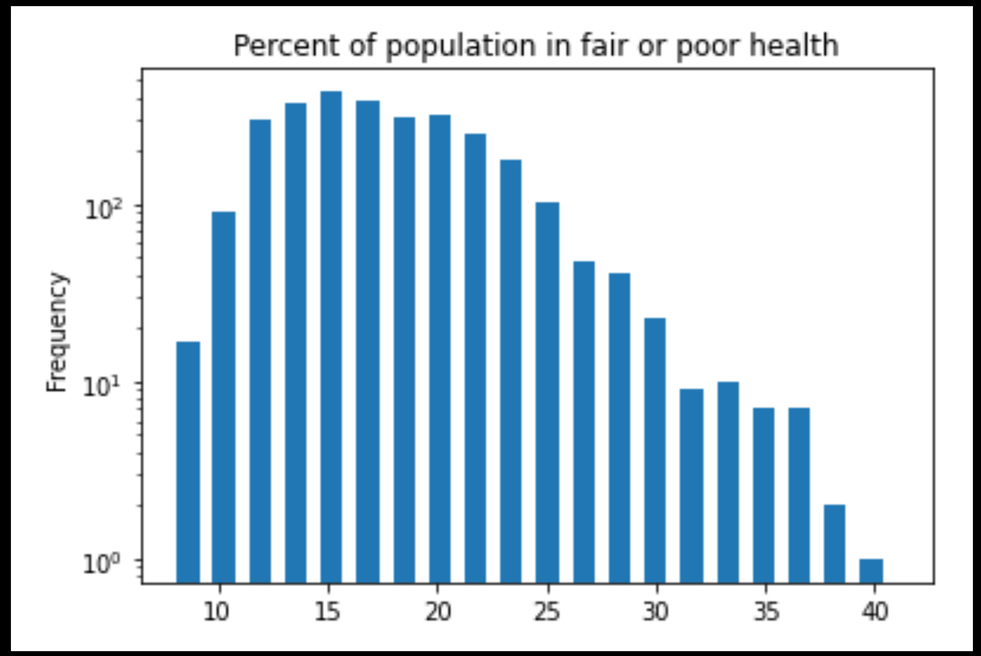


The major outlier at 200000 is New York County, one of the most affected areas in the country. The data is consistent with what is reported in other sources.



The same outlier exists in the deaths by county graph, once again New York County will not be removed from the data set as it is the true value.

Variable Histograms



The distribution of the graph somewhat resembles the sahpe of a normal distribution, no clear outliers exist.

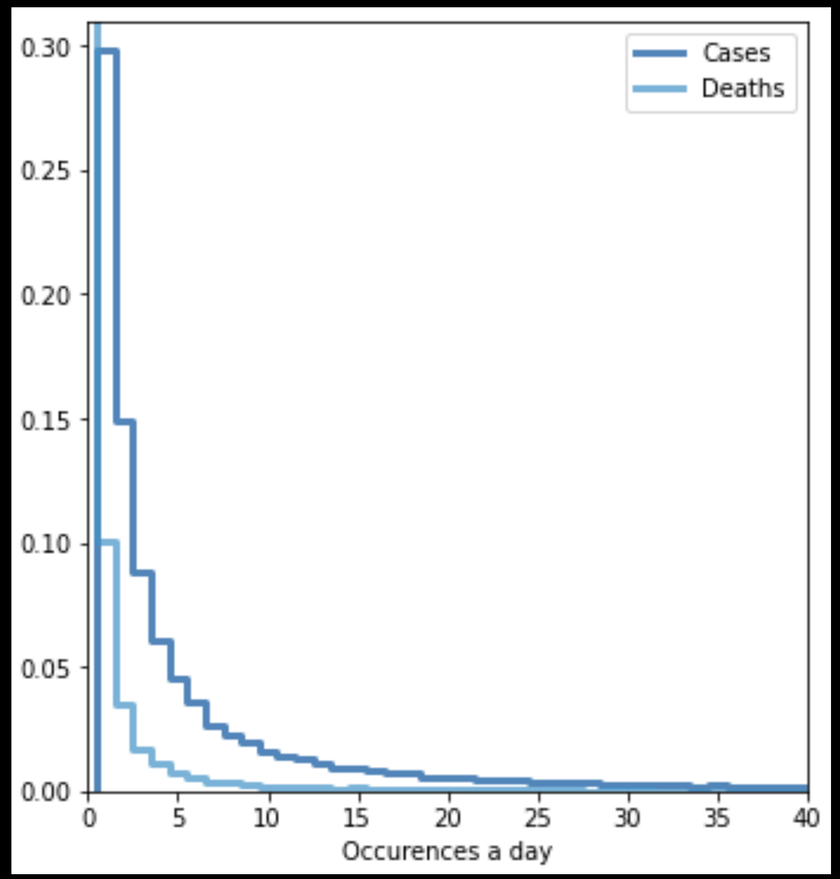
Variables - Descriptive Statistics

Both Data Sets

	mean	median	var	std
cases	581.075890	43.000000	2.089346e+07	4570.936959
deaths	34.123909	1.000000	1.686084e+05	410.619516
percent_fair_or_poor_health	18.034635	17.343802	2.234296e+01	4.726834
percent_smokers	17.532791	17.087545	1.255117e+01	3.542763
percent_adults_with_obesity	33.026591	33.300000	2.948427e+01	5.429942
percent_excessive_drinking	17.483325	17.559710	1.008055e+01	3.174989
income_ratio	4.520333	4.411360	5.491752e-01	0.741064
percent_adults_with_diabetes	12.237759	11.700000	1.635616e+01	4.044275

	mean	median	var	std
casenew	20.504134	3.0	14241.769449	119.338885
deathnew	1.165234	0.0	158.266400	12.580397

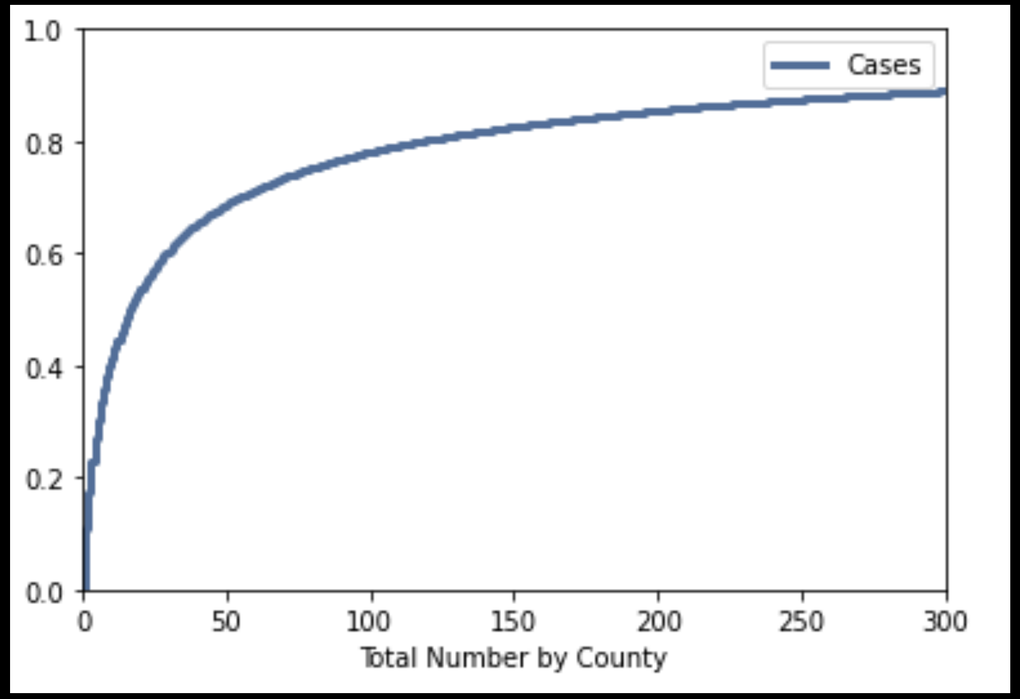
PMF Analysis



Initially was working with a filter to compare the data between the State of New York and the rest of the USA. Yet I realized that since Deaths are dependent on Cases, you can think of it as a subset with an unfortunate categorical variable.

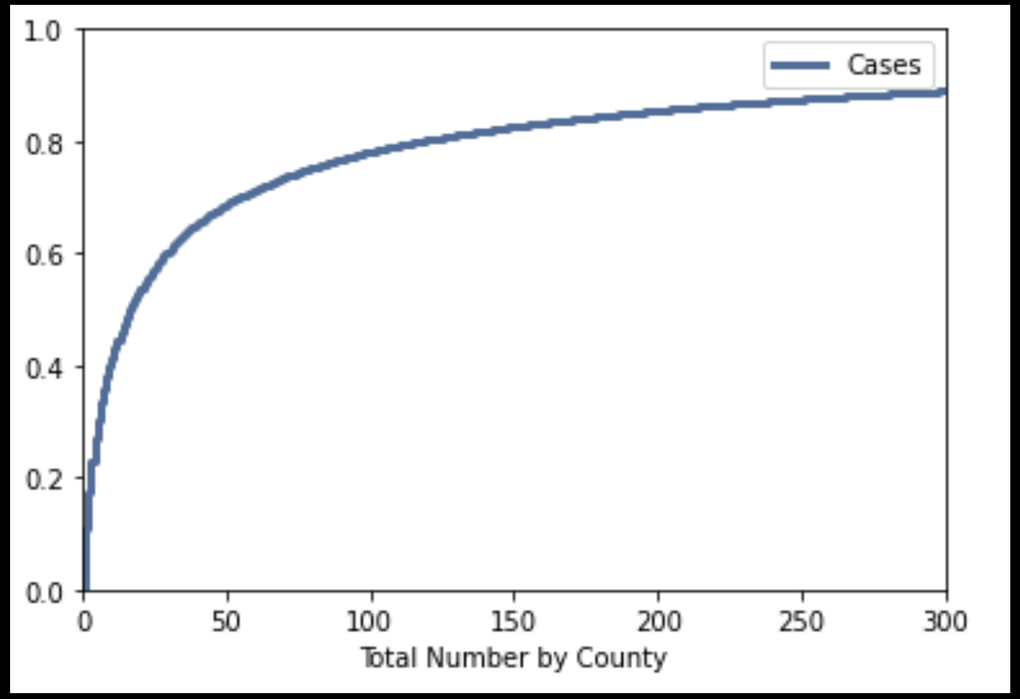
The data continues outside the scope of the graph but to be able to see the nuances I decided to cut off the tail end. The Light blue graph represents the probability of the number of people to die on a certain day due to coronavirus. The darker blue represents the probability of the number of cases that will happen on a certain day.

CDF Analysis



We can tell by the slope of the CDF that majority of the Counties (.7) are represented within 0-50 of cases. We also return to the outlier of New York at 200000. But once again to notice the proper curve of the data we have to cut off a certain amount of data. In the above graph we include about 90% of all cases, which tells us that 90% of all cases are in counties that have <300 cases in total.

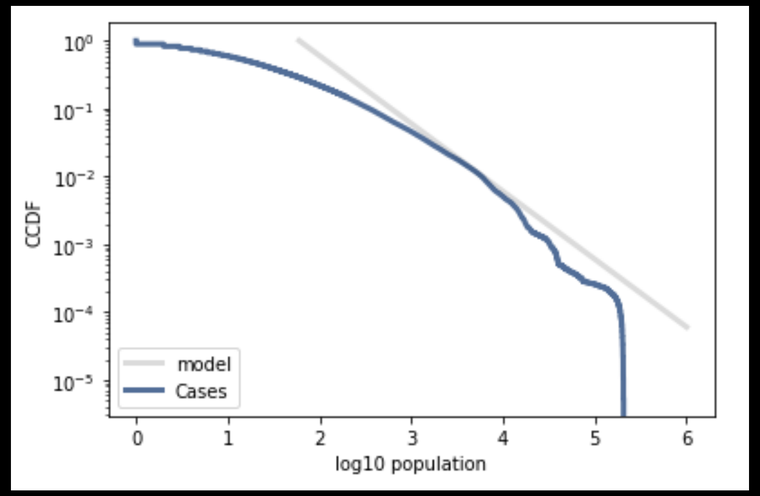
CDF Analysis



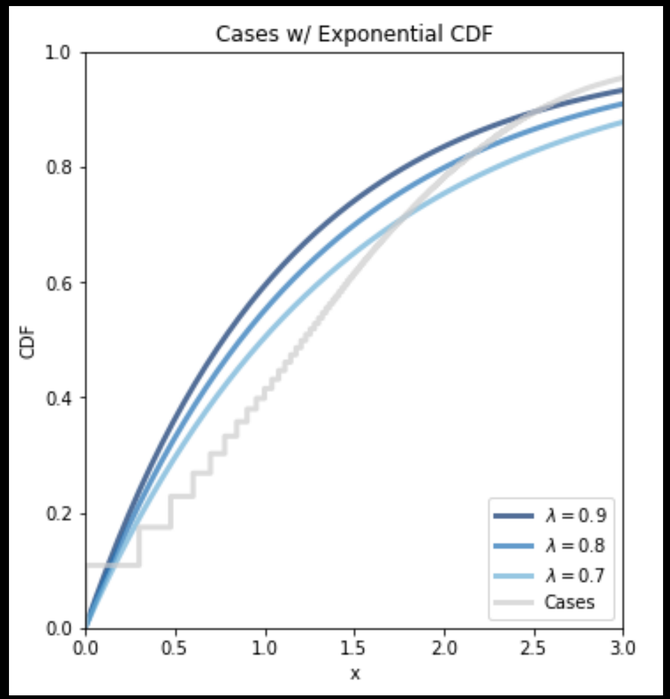
We can tell by the slope of the CDF that majority of the Counties (.7) are represented within 0-50 of cases. We also return to the outlier of New York at 200000. But once again to notice the proper curve of the data we have to cut off a certain amount of data. In the above graph we include about 90% of all cases, which tells us that 90% of all cases are in counties that have <300 cases in total.

Analytical Distribution

The CDF of the total number of cases resembles a Pareto Distribution. A key characteristic of a Pareto Dist. is that taking the log of each axis will produce a straight line.



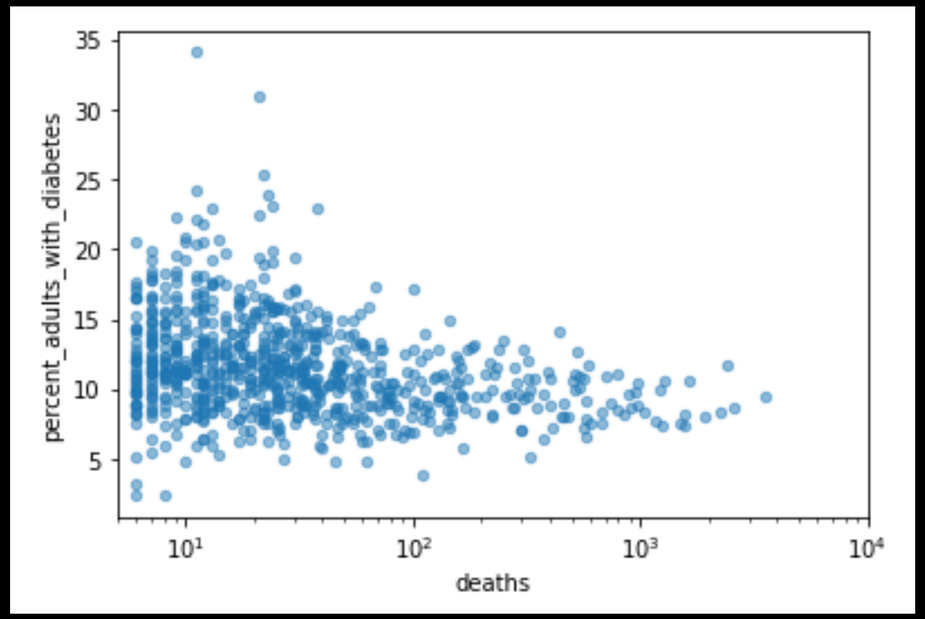
Turns out the data shouldn't be modeled by a Pareto Distribution, at least not the entire model. The middle section of the graph fits well to the data.



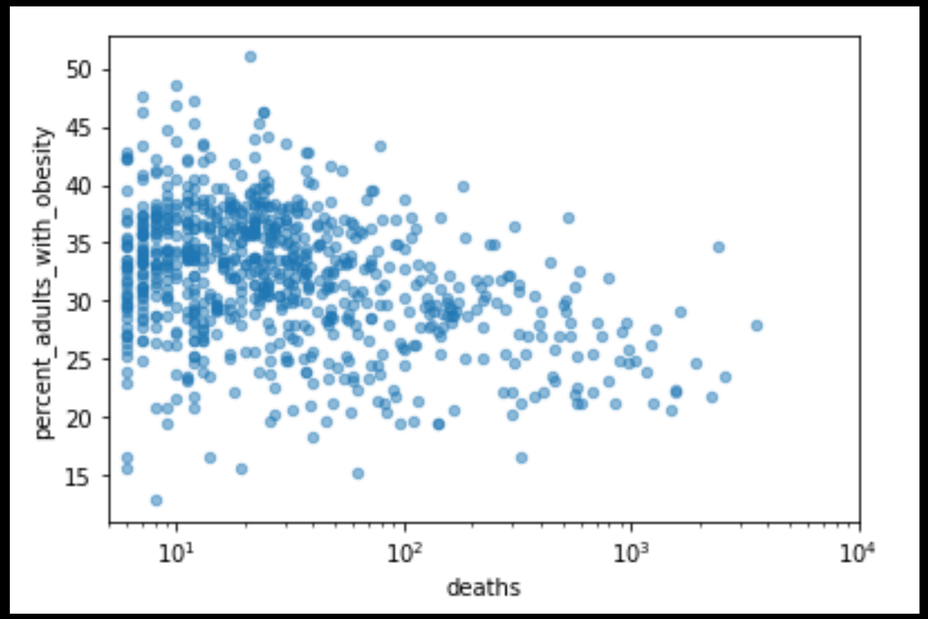
The Exponential Function is not a good estimate of the log-log CDF of Total Cases by County.

Analytical Distribution

Since some counties have luckily had no meaningful developments when it comes to cases or deaths, they end up being noise for the calculations since at the current time we are unable to clarify whether the counties have low numbers due to their health data or just because they have not yet been affected. These counties have been excluded and these are the resulting graphs. The correlations for the relationships among the variables was affected, strengthening the relationships. Below the graphs is the correlation value between the two variables.



-0.182626



-0.270842

Hypothesis Testing

As we are exploring the effects of comorbidities upon cases and deaths rates, we need to make sure that the relationships presented by the correlations are statistically sound. So we are going to be testing them to make sure of our hypothesis is confirmed.

Given the sample and the negative correlation observed in the relationship among deaths and percent adults with obesity what is the probability that this effect has occurred by chance?

```
In [520]: data = hatfilter.deaths, hatfilter.percent_adults_with_obesity
          ht = CorrelationPermute(data)
          pvalue = ht.PValue()
          pvalue

Out[520]: 0.0
```

The relationship we observe among the data is not by chance.

Regression Analysis

Total Deaths - Dependent Variable

OLS Regression Results

Dep. Variable:	deaths	**R-squared (uncentered):**	0.081
Model:	OLS	**Adj. R-squared (uncentered):**	0.080
Method:	Least Squares	**F-statistic:**	69.22
Date:	Sat, 30 May 2020	**Prob (F-statistic):**	3.89e-16
Time:	20:35:57	**Log-Likelihood:**	-5553.7
No. Observations:	786	**AIC:**	1.111e+04
Df Residuals:	785	**BIC:**	1.111e+04
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
percent_adults_with_obesity	2.5794	0.310	8.320	0.000	1.971	3.188

Omnibus:	966.975	**Durbin-Watson:**	1.927
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	91329.898
Skew:	6.295	**Prob(JB):**	0.00
Kurtosis:	54.286	**Cond. No.**	1.00

OLS Regression Results

Dep. Variable:	deaths	**R-squared (uncentered):**	0.083
Model:	OLS	**Adj. R-squared (uncentered):**	0.080
Method:	Least Squares	**F-statistic:**	35.27
Date:	Sat, 30 May 2020	**Prob (F-statistic):**	2.15e-15
Time:	20:37:56	**Log-Likelihood:**	-5553.1
No. Observations:	786	**AIC:**	1.111e+04
Df Residuals:	784	**BIC:**	1.112e+04
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
percent_adults_with_obesity	4.1786	1.435	2.911	0.004	1.361	6.996
percent_adults_with_diabetes	-4.4038	3.860	-1.141	0.254	-11.980	3.173

Omnibus:	967.428	**Durbin-Watson:**	1.921
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	91513.372
Skew:	6.300	**Prob(JB):**	0.00
Kurtosis:	54.338	**Cond. No.**	14.1

OLS Regression Results

Dep. Variable:	deaths	**R-squared (uncentered):**	0.082
Model:	OLS	**Adj. R-squared (uncentered):**	0.079
Method:	Least Squares	**F-statistic:**	34.80
Date:	Sat, 30 May 2020	**Prob (F-statistic):**	3.31e-15
Time:	20:37:31	**Log-Likelihood:**	-5553.5
No. Observations:	786	**AIC:**	1.111e+04
Df Residuals:	784	**BIC:**	1.112e+04
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
percent_adults_with_obesity	1.2392	2.056	0.603	0.547	-2.796	5.274
percent_smokers	2.5691	3.895	0.660	0.510	-5.077	10.215

Omnibus:	967.787	**Durbin-Watson:**	1.925
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	91746.810
Skew:	6.303	**Prob(JB):**	0.00
Kurtosis:	54.406	**Cond. No.**	16.0

OLS Regression Results

Dep. Variable:	deaths	**R-squared (uncentered):**	0.087
Model:	OLS	**Adj. R-squared (uncentered):**	0.085
Method:	Least Squares	**F-statistic:**	37.39
Date:	Sat, 30 May 2020	**Prob (F-statistic):**	3.08e-16
Time:	20:38:44	**Log-Likelihood:**	-5551.1
No. Observations:	786	**AIC:**	1.111e+04
Df Residuals:	784	**BIC:**	1.112e+04
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
percent_adults_with_obesity	-0.8618	1.540	-0.560	0.576	-3.885	2.161
percent_fair_or_poor_health	6.3026	2.763	2.281	0.023	0.879	11.726

Omnibus:	964.791	**Durbin-Watson:**	1.925
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	91146.118
Skew:	6.268	**Prob(JB):**	0.00
Kurtosis:	54.244	**Cond. No.**	11.6

After trying different combinations of predictive variables, only one of the models concluded with a solid significant p-value. That is the Simple Linear Regression, it describes that for every 1 percent in the county population that is obese, you can expect 2.6 deaths.

Thank you