Math 189Z Final Report

Anna Krutsinger, Keizo Morgan and Ingrid Tsang
https://github.com/amkrutsinger/covid19finalproject

Introduction

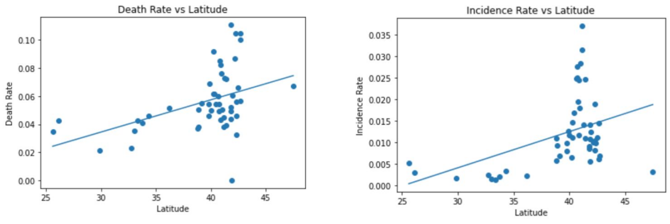
We sought to explore how geography impacts the incidence and morbidity of the novel coronavirus through investigating the correlations between various physical and human geographical variables and coronavirus incidence and death rates. US county-level data was used in this analysis as detailed data was available and counties represent a variety of geographies while within US country bounds, which allows for more consistency.

During the preliminary stages of the project, we considered using LDA to identify topic vectors related to geography from the cord19 library, a database of research papers from the COVID-19 Open Research Dataset Challenge. Upon crude implementation, however, performing LDA on both the article titles and abstracts did not yield fruitful results. Thus, we shifted towards a search-based approach, identifying pairs of geographical terms and features of COVID-19 to find key sentences in abstracts from a repository of scientific research. In doing so, we were able to parse relevant research and contextualize our project (see Appendix A).

With these geographical terms, we identified a number of geographical variables to perform linear regressions, including latitude, temperature, population density, and pollution levels. While there were fairly significant correlations in latitude, average temperature and average max heat index, supported by current literature, the correlations were not as strong with most of the other variables. Thus we attempted to create more predictive models using multivariate linear regressions through correlating multiple geographical variables with coronavirus incidence and death rates, which did not provide very conclusive findings due to an inability to process the data to satisfy all of the assumptions of OLS regression.

Linear Regression

The first step performed was linear regression, which was used to determine the correlations between single geographical variables and coronavirus incidence and death rates at the county level in the US. To make the data more reliable, only counties with more than 5000 cases were included in the analysis.



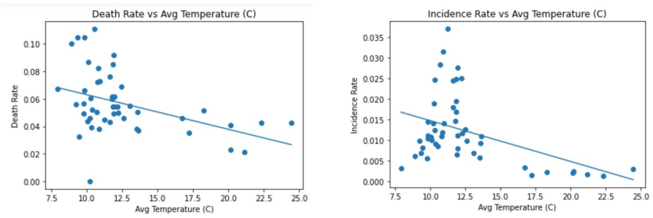
```
p-values: 0.0017079030043703537
R^2: 0.18706277936417032
Slope: 0.0023015954779224634
```

```
p-values: 0.0026187476210525256
R^2: 0.1735404140945842
Slope: 0.0008400356841260513
```

Figure 1: Linear regression of latitude against death rate and incidence rate.

Many geographical variables relating to both physical and human geography were analyzed. The variables with the most significant correlations with coronavirus incidence and death rates were latitude, average temperature, and average max heat index. These were determined based on the R^2 values (greater than 0.05) and p-values (less than 0.05 for statistical significance) of the linear regressions.

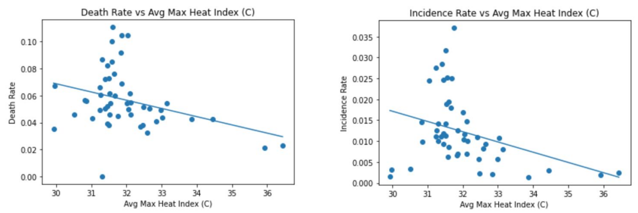
The two graphs in Figure 1 suggest that higher latitudes are correlated with higher coronavirus death and incidence rates. This is supported by studies[1] that have shown that cold and dry conditions boost the speed of transmission of coronavirus, since the climate is usually colder and drier at higher latitudes. These results are also corroborated by the graphs in Figures 2 and 3, which suggest lower average temperatures and max heat index (a measure of temperature by factoring in humidity) are correlated with higher death and incidence rates.



```
p-values: 0.002857323173418585
R^2: 0.17409602904666416
Slope: -0.0025174080196839342
```

```
p-values: 0.0019301302276176502
R^2: 0.18675282561657386
Slope: -0.0009907860465656498
```

Figure 2: Linear regression of average temperature against death rate and incidence rate.



```
p-values: 0.02050305832794961
R^2: 0.10902062715166252
Slope: -0.0061056782567494175
```

```
p-values: 0.014064474462451692
R^2: 0.12160207343116328
Slope: -0.0024504011036971187
```

Figure 3: Linear regression of average max heat index against death rate and incidence rate.

Multivariate Linear Regression

We used the Least Squares method to further investigate geographical variables found to have some correlation with the novel coronavirus' incidence and death rate. OLS assumes 5 characteristics: (1) the parameters of the linear regression model are linear, (2) errors are normally distributed, (3) the conditional mean is zero, (4) there is no perfect collinearity or multi-collinearity and (5) there is no autocorrelation and there is homoscedasticity.

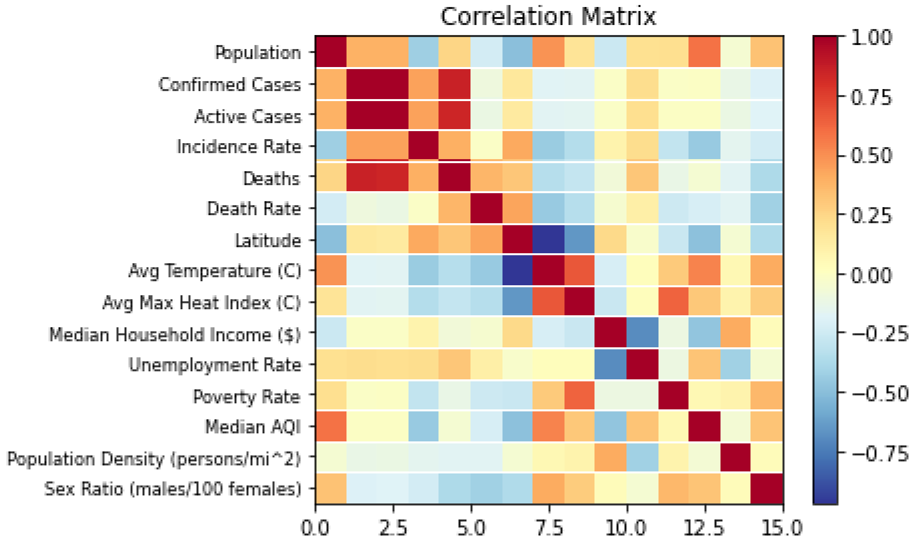
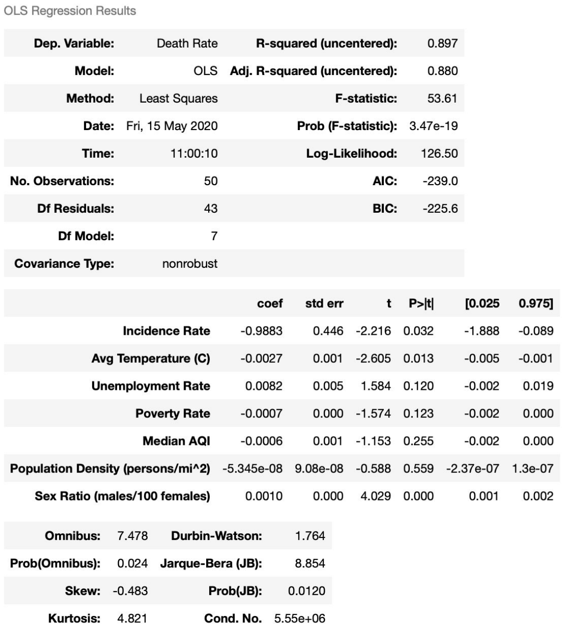


Figure 4: Correlation matrix between variables collected from counties with 5000 cases or greater.

[1] https://www.cebm.net/covid-19/do-weather-conditions-influence-the-transmission-of-the-coronavirus-sars-cov-2/

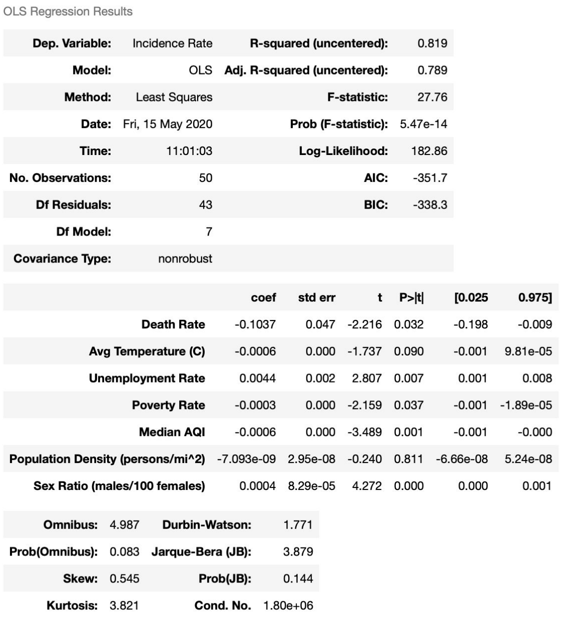


OLS Regression Results

Dep. Variable:	Death Rate	**R-squared (uncentered):**	0.897
Model:	OLS	**Adj. R-squared (uncentered):**	0.880
Method:	Least Squares	**F-statistic:**	53.61
Date:	Fri, 15 May 2020	**Prob (F-statistic):**	3.47e-19
Time:	11:00:10	**Log-Likelihood:**	126.50
No. Observations:	50	**AIC:**	-239.0
Df Residuals:	43	**BIC:**	-225.6
Df Model:	7		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
Incidence Rate	-0.9883	0.446	-2.216	0.032	-1.888	-0.089
Avg Temperature (C)	-0.0027	0.001	-2.605	0.013	-0.005	-0.001
Unemployment Rate	0.0082	0.005	1.584	0.120	-0.002	0.019
Poverty Rate	-0.0007	0.000	-1.574	0.123	-0.002	0.000
Median AQI	-0.0006	0.001	-1.153	0.255	-0.002	0.000
Population Density (persons/mi^2)	-5.345e-08	9.08e-08	-0.588	0.559	-2.37e-07	1.3e-07
Sex Ratio (males/100 females)	0.0010	0.000	4.029	0.000	0.001	0.002

Omnibus:	7.478	**Durbin-Watson:**	1.764
Prob(Omnibus):	0.024	**Jarque-Bera (JB):**	8.854
Skew:	-0.483	**Prob(JB):**	0.0120
Kurtosis:	4.821	**Cond. No.**	5.55e+06



OLS Regression Results

Dep. Variable:	Incidence Rate	**R-squared (uncentered):**	0.819
Model:	OLS	**Adj. R-squared (uncentered):**	0.789
Method:	Least Squares	**F-statistic:**	27.76
Date:	Fri, 15 May 2020	**Prob (F-statistic):**	5.47e-14
Time:	11:01:03	**Log-Likelihood:**	182.86
No. Observations:	50	**AIC:**	-351.7
Df Residuals:	43	**BIC:**	-338.3
Df Model:	7		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
Death Rate	-0.1037	0.047	-2.216	0.032	-0.198	-0.009
Avg Temperature (C)	-0.0006	0.000	-1.737	0.090	-0.001	9.81e-05
Unemployment Rate	0.0044	0.002	2.807	0.007	0.001	0.008
Poverty Rate	-0.0003	0.000	-2.159	0.037	-0.001	-1.89e-05
Median AQI	-0.0006	0.000	-3.489	0.001	-0.001	-0.000
Population Density (persons/mi^2)	-7.093e-09	2.95e-08	-0.240	0.811	-6.66e-08	5.24e-08
Sex Ratio (males/100 females)	0.0004	8.29e-05	4.272	0.000	0.000	0.001

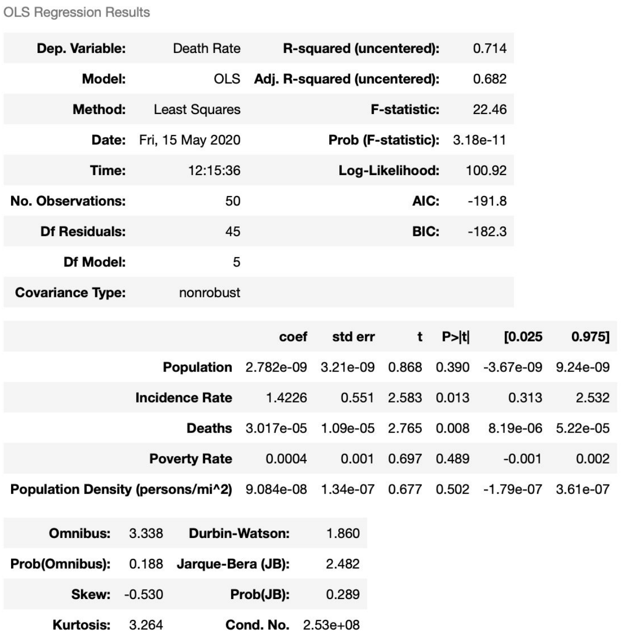
Omnibus:	4.987	**Durbin-Watson:**	1.771
Prob(Omnibus):	0.083	**Jarque-Bera (JB):**	3.879
Skew:	0.545	**Prob(JB):**	0.144
Kurtosis:	3.821	**Cond. No.**	1.80e+06

Figure 5: Summary of outputs for multiple linear regressions, death rate (left) and incidence rate (right)

The first OLS regression treated death rate as a dependent variable, and incidence rate, average temperature (C), unemployment rate, poverty rate, median AQI, population density (persons per square mile), and sex ratio as independent variables. The R^2 value (0.897) indicates that 89.7% variation in the death rate can be explained by the independent variables. However, it is important to note that the statistic increases when the number of predictors increases, so it is inconclusive as to whether adding or omitting certain variables actually increases how powerful the predictions of the regression are. Since the Prob(F-statistic) is close to zero (3.47e-19), this suggests that overall, our regression is meaningful. Furthermore, from the Durbin Watson test statistic (1.764), we see that the variance of errors is relatively consistent across the dataset since it is between 1 and 2, satisfying the OLS homoscedasticity assumption. This suggests that our results are reliable from an interpretative standpoint. However, the model does not satisfy some of the key assumptions listed previously. For one, the results from the Omnibus Test indicate that errors are not normally distributed, as the probability that the residuals are distributed normally is 0.024. This means that the estimated coefficients are not Best Linear Unbiased Estimators as desired. Furthermore the large condition number (5.55e+06) suggests that there are strong multicollinearity or other numerical problems.

The second OLS regression treated incidence rate as a dependent variable, and death rate, average temperature (C), unemployment rate, poverty rate, median AQI, population density (persons per square mile), and sex ratio as independent variables. The R^2 value (0.789) indicates that 78.9% variation in incidence rate can be explained by the independent variables. A Prob(F-statistic) close to zero (5.47e-14) suggests that our regression is meaningful. Furthermore, from the Durbin Watson test statistic (1.771), we see that the variance of errors is relatively consistent across the dataset since it is between 1 and 2, so our results can be reliable from an interpretative standpoint. However, like the first OLS regression, this model does not satisfy some of the key assumptions listed previously. For one, the probability that the residuals are distributed normally is 0.083, meaning the estimated coefficients are not Best Linear Unbiased Estimators as desired. Furthermore the large condition number (1.80e+06) suggests that there are strong multicollinearity or other numerical problems.

To address the multicollinearity issue, we removed highly correlated predictors from the model that had a variance inflation factor greater than 2.5, and re-ran the regressions. This yielded the outputs below. These results indicated that strong multicollinearity cannot wholly explain the large condition number.

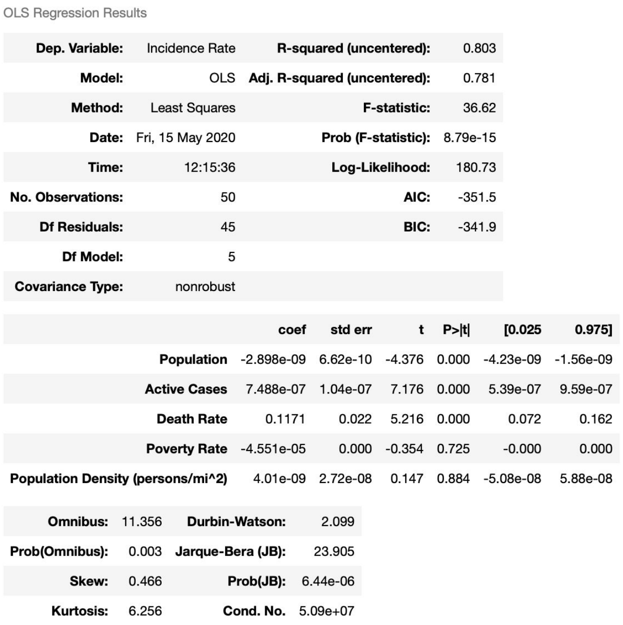


OLS Regression Results

Dep. Variable:	Death Rate	**R-squared (uncentered):**	0.714
Model:	OLS	**Adj. R-squared (uncentered):**	0.682
Method:	Least Squares	**F-statistic:**	22.46
Date:	Fri, 15 May 2020	**Prob (F-statistic):**	3.18e-11
Time:	12:15:36	**Log-Likelihood:**	100.92
No. Observations:	50	**AIC:**	-191.8
Df Residuals:	45	**BIC:**	-182.3
Df Model:	5		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
Population	2.782e-09	3.21e-09	0.868	0.390	-3.67e-09	9.24e-09
Incidence Rate	1.4226	0.551	2.583	0.013	0.313	2.532
Deaths	3.017e-05	1.09e-05	2.765	0.008	8.19e-06	5.22e-05
Poverty Rate	0.0004	0.001	0.697	0.489	-0.001	0.002
Population Density (persons/mi^2)	9.084e-08	1.34e-07	0.677	0.502	-1.79e-07	3.61e-07

Omnibus:	3.338	**Durbin-Watson:**	1.860
Prob(Omnibus):	0.188	**Jarque-Bera (JB):**	2.482
Skew:	-0.530	**Prob(JB):**	0.289
Kurtosis:	3.264	**Cond. No.**	2.53e+08



OLS Regression Results

Dep. Variable:	Incidence Rate	**R-squared (uncentered):**	0.803
Model:	OLS	**Adj. R-squared (uncentered):**	0.781
Method:	Least Squares	**F-statistic:**	36.62
Date:	Fri, 15 May 2020	**Prob (F-statistic):**	8.79e-15
Time:	12:15:36	**Log-Likelihood:**	180.73
No. Observations:	50	**AIC:**	-351.5
Df Residuals:	45	**BIC:**	-341.9
Df Model:	5		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
Population	-2.898e-09	6.62e-10	-4.376	0.000	-4.23e-09	-1.56e-09
Active Cases	7.488e-07	1.04e-07	7.176	0.000	5.39e-07	9.59e-07
Death Rate	0.1171	0.022	5.216	0.000	0.072	0.162
Poverty Rate	-4.551e-05	0.000	-0.354	0.725	-0.000	0.000
Population Density (persons/mi^2)	4.01e-09	2.72e-08	0.147	0.884	-5.08e-08	5.88e-08

Omnibus:	11.356	**Durbin-Watson:**	2.099
Prob(Omnibus):	0.003	**Jarque-Bera (JB):**	23.905
Skew:	0.466	**Prob(JB):**	6.44e-06
Kurtosis:	6.256	**Cond. No.**	5.09e+07

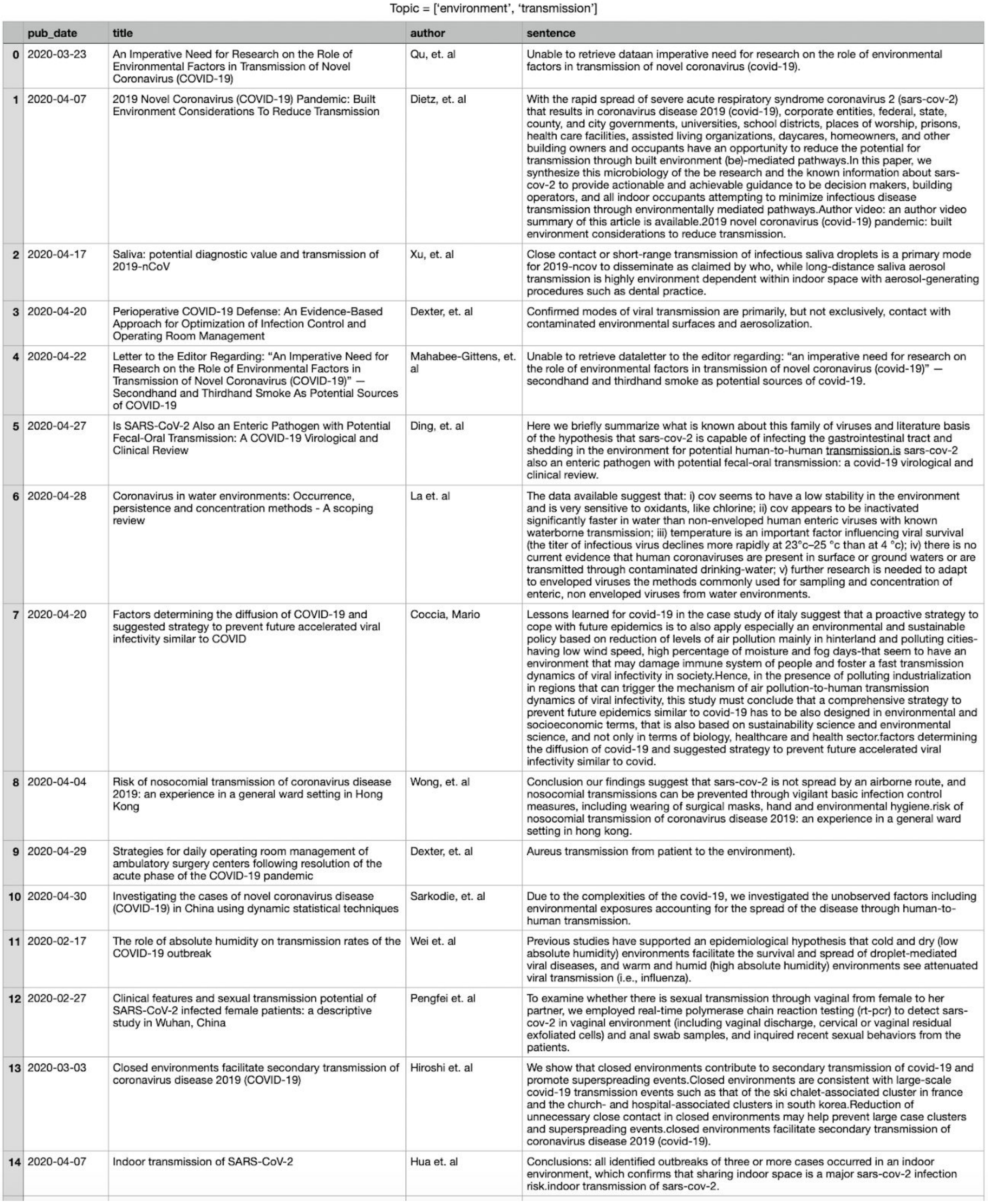
Figure 6: Summary of outputs for multiple linear regressions after VIF filtering, death rate (left) and incidence rate (right)

Conclusion

Our model sought out to find the relationship between the death and incidence rates with geographic factors such as temperature or population density. We searched through research papers in order to find factors of interest, and then amalgamated different datasets by county. The multiple linear regressions we ran suggest a strong relationship between these geographic factors with the death and incidence rates, given the R^2 values of .897 and .819, respectively, however after checking the assumptions for out multiple linear regression such as for multicollinearity, we are not entirely confident these relationships are linear.

In addition to investigating multicollinearity, another potential issue with our model may have been nonlinear relationships between our variables. We ran Q-Q plots and noted many of the variables used were heavily skewed (see Appendix B). For example: population, where the highest percentile of counties by county have drastically larger populations than the rest of the samples. Some form of transformation such as a log-transformation may have been required to create linearity, however we did not have the time to investigate how to proceed. Thus, if we were to continue this investigation, we would find a strategy to address the skew and heavy-tailedness of our input variables in order to make our model more predictive.

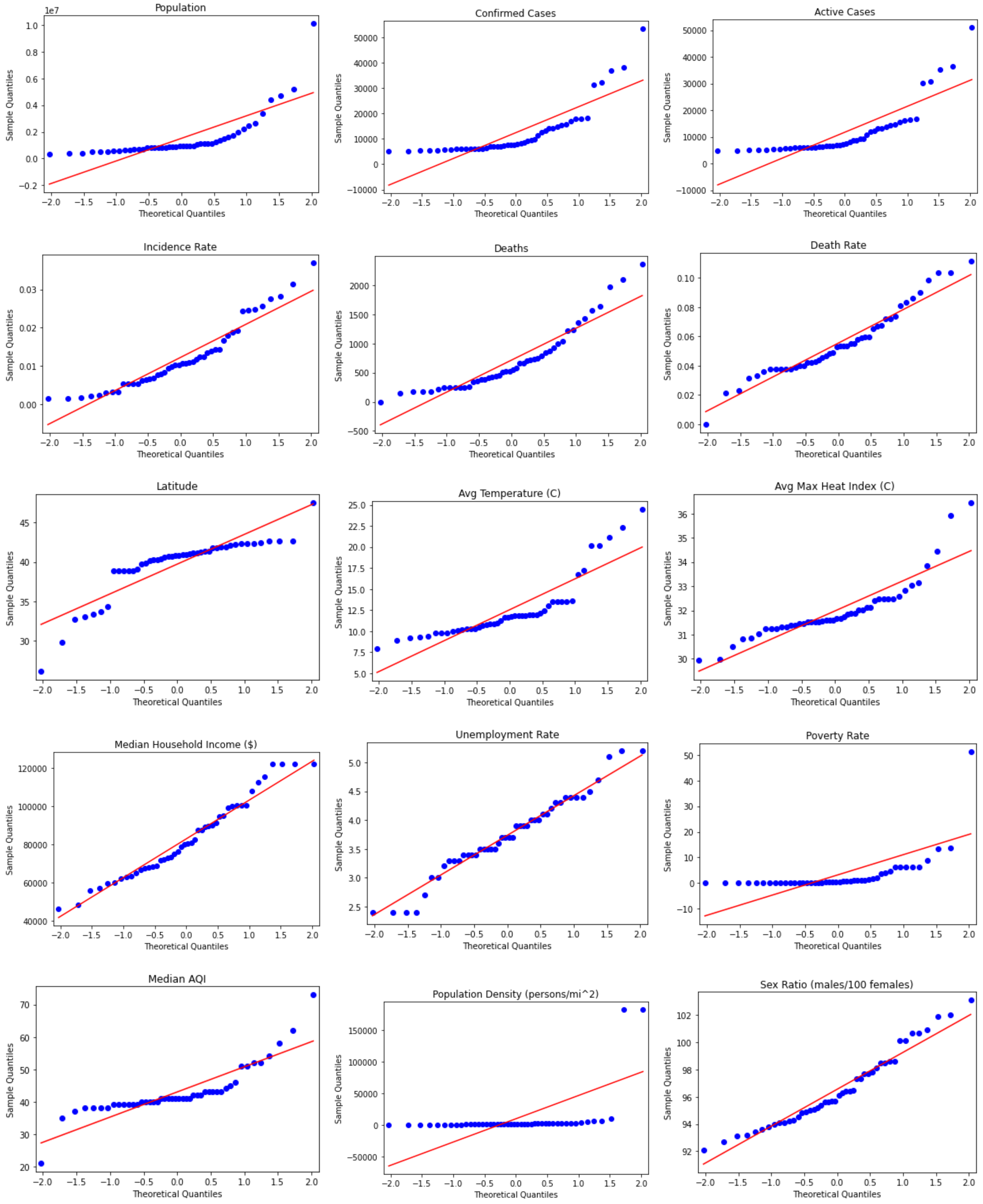
Appendix A: Topic Parsing



Topic = ['environment', 'transmission']

	pub_date	title	author	sentence
0	2020-03-23	An Imperative Need for Research on the Role of Environmental Factors in Transmission of Novel Coronavirus (COVID-19)	Qu, et. al	Unable to retrieve dataan imperative need for research on the role of environmental factors in transmission of novel coronavirus (covid-19).
1	2020-04-07	2019 Novel Coronavirus (COVID-19) Pandemic: Built Environment Considerations To Reduce Transmission	Dietz, et. al	With the rapid spread of severe acute respiratory syndrome coronavirus 2 (sars-cov-2) that results in coronavirus disease 2019 (covid-19), corporate entities, federal, state, county, and city governments, universities, school districts, places of worship, prisons, health care facilities, assisted living organizations, daycares, homeowners, and other building owners and occupants have an opportunity to reduce the potential for transmission through built environment (be)-mediated pathways.In this paper, we synthesize this microbiology of the be research and the known information about sars-cov-2 to provide actionable and achievable guidance to be decision makers, building operators, and all indoor occupants attempting to minimize infectious disease transmission through environmentally mediated pathways.Author video: an author video summary of this article is available.2019 novel coronavirus (covid-19) pandemic: built environment considerations to reduce transmission.
2	2020-04-17	Saliva: potential diagnostic value and transmission of 2019-nCoV	Xu, et. al	Close contact or short-range transmission of infectious saliva droplets is a primary mode for 2019-ncov to disseminate as claimed by who, while long-distance saliva aerosol transmission is highly environment dependent within indoor space with aerosol-generating procedures such as dental practice.
3	2020-04-20	Perioperative COVID-19 Defense: An Evidence-Based Approach for Optimization of Infection Control and Operating Room Management	Dexter, et. al	Confirmed modes of viral transmission are primarily, but not exclusively, contact with contaminated environmental surfaces and aerosolization.
4	2020-04-22	Letter to the Editor Regarding: "An Imperative Need for Research on the Role of Environmental Factors in Transmission of Novel Coronavirus (COVID-19)" — Secondhand and Thirdhand Smoke As Potential Sources of COVID-19	Mahabee-Gittens, et. al	Unable to retrieve dataletter to the editor regarding: "an imperative need for research on the role of environmental factors in transmission of novel coronavirus (covid-19)" — secondhand and thirdhand smoke as potential sources of covid-19.
5	2020-04-27	Is SARS-CoV-2 Also an Enteric Pathogen with Potential Fecal-Oral Transmission: A COVID-19 Virological and Clinical Review	Ding, et. al	Here we briefly summarize what is known about this family of viruses and literature basis of the hypothesis that sars-cov-2 is capable of infecting the gastrointestinal tract and shedding in the environment for potential human-to-human transmission.is sars-cov-2 also an enteric pathogen with potential fecal-oral transmission: a covid-19 virological and clinical review.
6	2020-04-28	Coronavirus in water environments: Occurrence, persistence and concentration methods - A scoping review	La et. al	The data available suggest that: i) cov seems to have a low stability in the environment and is very sensitive to oxidants, like chlorine; ii) cov appears to be inactivated significantly faster in water than non-enveloped human enteric viruses with known waterborne transmission; iii) temperature is an important factor influencing viral survival (the titer of infectious virus declines more rapidly at 23°c–25 °c than at 4 °c); iv) there is no current evidence that human coronaviruses are present in surface or ground waters or are transmitted through contaminated drinking-water; v) further research is needed to adapt to enveloped viruses the methods commonly used for sampling and concentration of enteric, non enveloped viruses from water environments.
7	2020-04-20	Factors determining the diffusion of COVID-19 and suggested strategy to prevent future accelerated viral infectivity similar to COVID	Coccia, Mario	Lessons learned for covid-19 in the case study of italy suggest that a proactive strategy to cope with future epidemics is to also apply especially an environmental and sustainable policy based on reduction of levels of air pollution mainly in hinterland and polluting cities-having low wind speed, high percentage of moisture and fog days-that seem to have an environment that may damage immune system of people and foster a fast transmission dynamics of viral infectivity in society.Hence, in the presence of polluting industrialization in regions that can trigger the mechanism of air pollution-to-human transmission dynamics of viral infectivity, this study must conclude that a comprehensive strategy to prevent future epidemics similar to covid-19 has to be also designed in environmental and socioeconomic terms, that is also based on sustainability science and environmental science, and not only in terms of biology, healthcare and health sector.factors determining the diffusion of covid-19 and suggested strategy to prevent future accelerated viral infectivity similar to covid.
8	2020-04-04	Risk of nosocomial transmission of coronavirus disease 2019: an experience in a general ward setting in Hong Kong	Wong, et. al	Conclusion our findings suggest that sars-cov-2 is not spread by an airborne route, and nosocomial transmissions can be prevented through vigilant basic infection control measures, including wearing of surgical masks, hand and environmental hygiene.risk of nosocomial transmission of coronavirus disease 2019: an experience in a general ward setting in hong kong.
9	2020-04-29	Strategies for daily operating room management of ambulatory surgery centers following resolution of the acute phase of the COVID-19 pandemic	Dexter, et. al	Aureus transmission from patient to the environment).
10	2020-04-30	Investigating the cases of novel coronavirus disease (COVID-19) in China using dynamic statistical techniques	Sarkodie, et. al	Due to the complexities of the covid-19, we investigated the unobserved factors including environmental exposures accounting for the spread of the disease through human-to-human transmission.
11	2020-02-17	The role of absolute humidity on transmission rates of the COVID-19 outbreak	Wei et. al	Previous studies have supported an epidemiological hypothesis that cold and dry (low absolute humidity) environments facilitate the survival and spread of droplet-mediated viral diseases, and warm and humid (high absolute humidity) environments see attenuated viral transmission (i.e., influenza).
12	2020-02-27	Clinical features and sexual transmission potential of SARS-CoV-2 infected female patients: a descriptive study in Wuhan, China	Pengfei et. al	To examine whether there is sexual transmission through vaginal from female to her partner, we employed real-time polymerase chain reaction testing (rt-pcr) to detect sars-cov-2 in vaginal environment (including vaginal discharge, cervical or vaginal residual exfoliated cells) and anal swab samples, and inquired recent sexual behaviors from the patients.
13	2020-03-03	Closed environments facilitate secondary transmission of coronavirus disease 2019 (COVID-19)	Hiroshi et. al	We show that closed environments contribute to secondary transmission of covid-19 and promote superspreading events.Closed environments are consistent with large-scale covid-19 transmission events such as that of the ski chalet-associated cluster in france and the church- and hospital-associated clusters in south korea.Reduction of unnecessary close contact in closed environments may help prevent large case clusters and superspreading events.closed environments facilitate secondary transmission of coronavirus disease 2019 (covid-19).
14	2020-04-07	Indoor transmission of SARS-CoV-2	Hua et. al	Conclusions: all identified outbreaks of three or more cases occurred in an indoor environment, which confirms that sharing indoor space is a major sars-cov-2 infection risk.indoor transmission of sars-cov-2.

Appendix B: Quantile–Quantile Plots of All Variables


Population
1e7
Confirmed Cases
Active Cases
Incidence Rate
Deaths
Death Rate
Latitude
Avg Temperature (C)
Avg Max Heat Index (C)
Median Household Income ($)
Unemployment Rate
Poverty Rate
Median AQI
Population Density (persons/mi^2)
Sex Ratio (males/100 females)
Sample Quantiles
Theoretical Quantiles