Project Milestone 2:

Data Analysis on COVID-19 Cases in Italy

Group 57

Elizabeth Lentine, Rohan Bansal, Drew Williams, Divya Ramakrishnan

Objectives

We decided to analyze a dataset containing information regarding the spread of Coronavirus in Italy. The justification for choosing this topic is that we wanted to explore a topic for this project that would provide interesting and relevant information on a matter that concerns us all. We felt that there was no better topic to explore than the COVID-19 pandemic.

The Italy dataset we found, sourced from the Italian government, contains 19 features [A.0] and is updated several times a day. This data includes information on the country, each province, and each region. Since it provided so many useful measures and we can count on current information being added, we felt this was a good dataset to answer some of our questions about Italy. We will explore how social distancing and testing measures have affected Italy, and how the spread has been contained from the initial outbreak to now. Furthermore, we hope to see how effective the social distancing measures were on the rate of infection. Lastly, we would like to see if there were any surprising spikes shown from the dataset in regards to the spread of the virus. If there were random spikes in the number of infections, despite strict distancing measures in place, it would be interesting to further examine that.

Approach #1: Linear Regression

[Figure 1] Linear model to predict death rate based on total cases

Dep. Variable:	deceduti	**R-squared (uncentered):**	0.941
Model:	OLS	**Adj. R-squared (uncentered):**	0.941
Method:	Least Squares	**F-statistic:**	1204.
Date:	Sun, 10 May 2020	**Prob (F-statistic):**	6.01e-48
Time:	15:27:19	**Log-Likelihood:**	-742.30
No. Observations:	76	**AIC:**	1487.
Df Residuals:	75	**BIC:**	1489.
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
totale_positivi	0.2234	0.006	34.703	0.000	0.211	0.236

Omnibus:	15.235	**Durbin-Watson:**	0.010
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	17.204
Skew:	1.117	**Prob(JB):**	0.000184
Kurtosis:	3.663	**Cond. No.**	1.00

We created a linear model using data on the number of deaths and total cases in all of Italy to predict the number of deaths based on the amount of the people who have tested positive for COVID-19. The model format is as follows:

$$Y = ßX + E$$

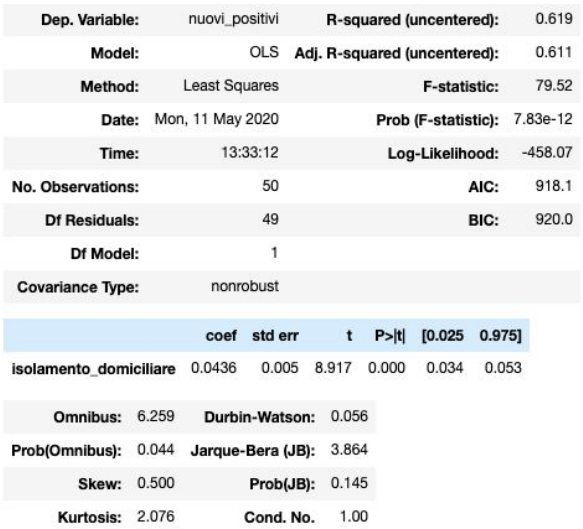
where Y corresponds to deaths, and X is total COVID-19 cases in Italy. The results show us that for an increase in the number of total positive cases there is a .2234 increase in the amount of deaths. This coefficient is statistically significant because the p-value associated with it is essentially zero. Additionally, the R-Squared value is close to 1, so the data is nicely fitted to the regression line. The coefficient value is somewhat surprising, but it is likely explained by the fact that Italy was one of the first countries hit by COVID-19 and it is likely that their treatment for the virus at first was worse than other countries who got hit by the pandemic a little bit later.

We created a linear model to predict the number of new cases in Italy based on the number of people in isolation according to the model:

$$Y = ßX + E$$

where Y is the number of new cases and X is the number of people in quarantine. Before performing this linear model, we predicted that there would be a negative coefficient, speculating that more people in isolation would lead to the virus spreading less. This difference is likely due to the fact that at the beginning when people were first isolating there were a lot of cases of COVID-19, so it did not directly result in less new cases at first.

[Figure 2] Linear Model to Predict New Cases by People in Isolation:



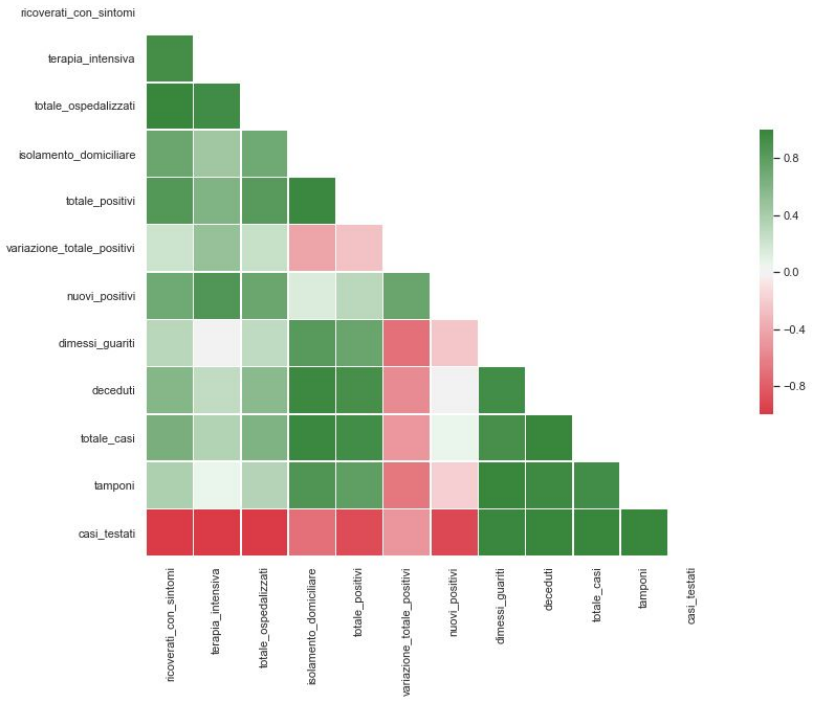
Dep. Variable:	nuovi_positivi	**R-squared (uncentered):**	0.619
Model:	OLS	**Adj. R-squared (uncentered):**	0.611
Method:	Least Squares	**F-statistic:**	79.52
Date:	Mon, 11 May 2020	**Prob (F-statistic):**	7.83e-12
Time:	13:33:12	**Log-Likelihood:**	-458.07
No. Observations:	50	**AIC:**	918.1
Df Residuals:	49	**BIC:**	920.0
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
isolamento_domiciliare	0.0436	0.005	8.917	0.000	0.034	0.053

Omnibus:	6.259	**Durbin-Watson:**	0.056
Prob(Omnibus):	0.044	**Jarque-Bera (JB):**	3.864
Skew:	0.500	**Prob(JB):**	0.145
Kurtosis:	2.076	**Cond. No.**	1.00

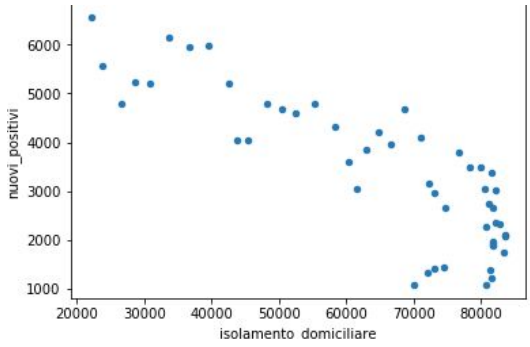
We chose to investigate “isolamento domiciliare”, or the number of people in isolation, to assess the effect of social distancing. Of the features, this is the only one involving a behavioral or regulatory component. The heat map below displays the correlations between features. To measure the efficacy of social distancing, we further consider the entries relating to isolation. From the graphic, some notable relationships are the positive correlations between isolation and total cases, deaths, and current positive cases. This makes intuitive sense for Italy, because from February until quite recently, the number of cases and deaths were still increasing and there was still a collaborative effort to combat those increases by staying inside.

[Figure 3] Correlation Heat Map of Italy COVID-19 Dataset Features



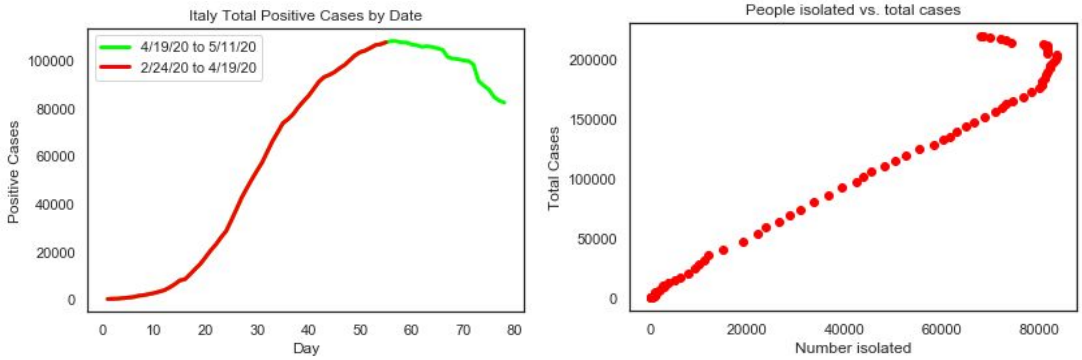
However, when we plot the number of people isolated vs. new positive results, we see a negatively-sloped trend as below:

[Figure 4] People Isolated vs. New Positive Cases



Furthermore, this model's R-squared value for the linear regression is lower, indicating that the data is not that closely fitted to the regression line. Taking a closer look at the isolation figures, the maximum number of people in isolation was on day 65, or April 28th. By April 29th, the isolation numbers were declining. Since Italy has passed the maximum point of the curve, it is important to keep in mind how the data will be affected. It is interesting to note how differently our data is distributed now in comparison to Milestone I of this report, as it seems that Italy was at the maximum of the curve at that time.

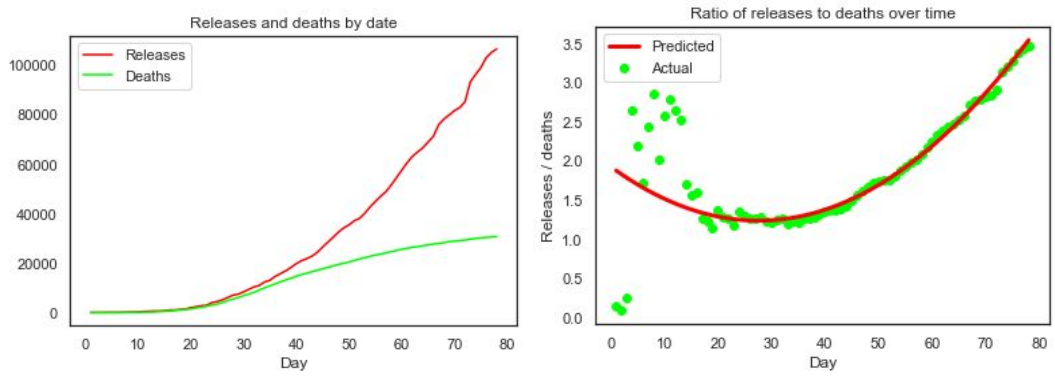
[Figures 5 and 6] Italy's COVID-19 Curve and the Influence of Isolation



After April 28th, our data shows curves that have doubled back in accordance with the fact that Italians were beginning to come out of isolation. The slopes of the second curve in figures 4 and 6 show that even as this happens, the new cases are much lower than they were at the same isolation levels not much earlier. This could be explained in a number of ways - namely the possibility of growing herd immunity[A.1] and Italians' adherence to social distancing guidelines. This change in the data illustrates the difficulty in using a strictly linear model to predict total cases based on isolation.

Approach #2: Polynomial Fit

[Figures 7 and 8] Comparing hospital releases to deaths over time



Here we can see that the hospitals started out with each COVID patient they received ultimately dying. After day ~20, there is a convex-parabolically shaped increase in the ratio of releases to deaths. We fit a polynomial with degree = 2 to the ratio of releases to deaths and find that this model actually predicts the rate quite well from day 20 on. Clearly there is more variance in the residuals during the first few weeks, as the number of patients was much smaller and the rates varied drastically by day during that time period.

This shape might signal that the hospitals have become better prepared to treat COVID patients over time. The behavior of the release to death ratio at the beginning stands out as unusual. There were not as many COVID patients in that time, and it is likely that as the disease spread, the healthcare system was overwhelmed and not able to provide life-saving treatment to everyone who was admitted. After day 20, this rate is clearly improving. This could be attributed to the arrival of new equipment, foreign doctors, or that the first patients to seek treatment were likely in incredibly critical condition compared to future infected individuals.

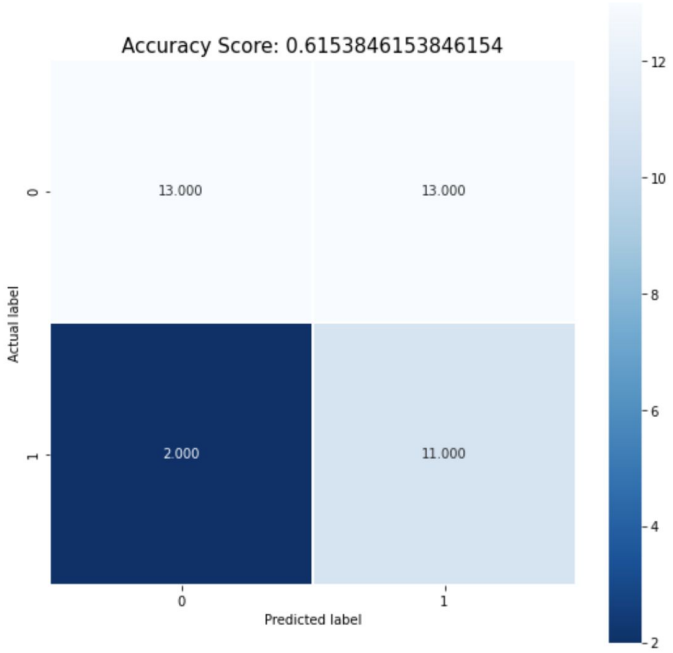
Approach #3: Logistic Regression

We ran a logistic regression to see if there could be an accurate model to predict whether the number of positive cases of COVID-19 on the current day was greater than the number of positive cases on the day before. One of our original questions revolved around seeing if the mitigation measures actually worked in slowing the number of positive cases in Italy, and I thought that creating a logistic regression around it would reveal interesting information. The dependent variable, **y**, is defined as follows:

0 - positive COVID-19 cases on the current day ≯ positive cases on the previous day
1 - positive COVID-19 cases on the current day **>** number of positive COVID-19 cases on the previous day.

The independent variables were different containment measures in Italy. These were "number of people in home isolation", "number of people tested", and "number of people hospitalized". The test_train_split method was used to create a 50/50 split. After creating the model, it displayed an accuracy score of .61538. We found this relatively acceptable given that this data has only been collected over less than three months. Additionally, the lower accuracy score could represent human fallacy in the containment measures on certain days - leading to unpredictable increases or decreases in total positive cases.

[Figure 9] Confusion Matrix of Logistic Regression as a Heat Map

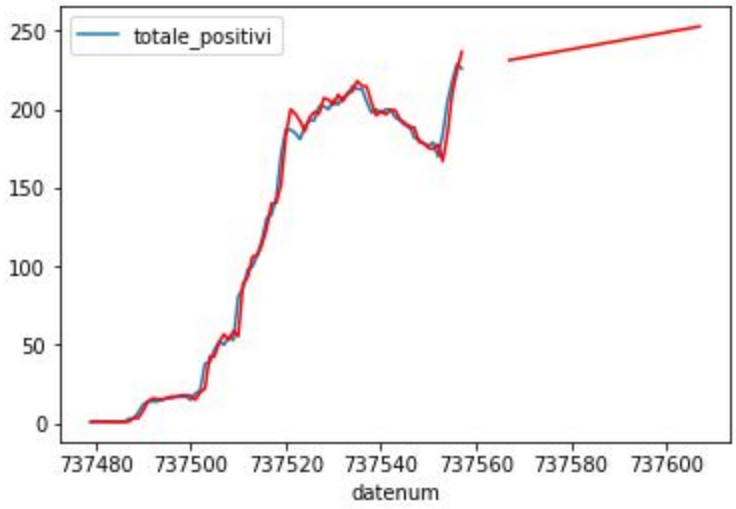


The confusion matrix revealed that the model predicted a large number of true positives correctly, as shown that FP = 2 and TP = 11. That indicates that it was able to yield decently accurate predictions of when the number of positive cases was greater on the current day than the previous. Overall, the data needs to be collected over a longer period of time, and the covariance between the independent variables should be measured to see if they are truly independent. However, based on the logistic regression model that was created, it seems that there is a statistical trend regarding the number of positive cases on a given day in the dataset compared to the day before it.

Approach #4: Forecasting

We used an Exponential Smoothing model to forecast the direction of the total number of positive cases for each region. For this model, we chose to use an additive trend model, chosen according to its lower sum of squared errors (3531.59 as compared to 4332.19 for a multiplicative model). By analyzing the resulting graph for each region using the data available since February 24th 2020, we were able to visualize the predicted trend for these positive cases. Out of the 20 regions of Italy, the model predicted 19 regions to decrease in positive cases in the near future given the time series data. Notably, the model predicted a significant increase in total positive cases for Molise. This makes for an interesting outlier to investigate and compare with the other regions.

[Figure 10] Forecasted Total Positive Cases in Molise



Here, the red line shows the fit of the Exponential Smoothing model to the data for total positive cases, and the following red line shows the forecasted trend in total positive cases in the future. The predicted increase in cases could imply that the current containment methods are not working as well as the rest of Italy. However, by checking the data, we can see that Molise has only had 386 confirmed total cases. It is likely that Molise is much earlier in the stages of progression of spread of coronavirus than the other regions.

Conclusion

The tools we used in this analysis provided us with numerous insights into the effects of COVID-19 on Italy. We were able to glean much information about the spread of the virus in this region, but we also realized that due to the short time frame of the length of

this pandemic, the dataset was small and could have produced inaccuracies. Our results are also subject to overfitting to the conditions of Italy and would be difficult to extrapolate to another country or pandemic.

Appendix

[A.0] Italy dataset: features

1. Date
2. Country
3. Region code
4. Region name
5. Latitude
6. Longitude
7. Hospitalized with symptoms
8. Intensive care
9. Total hospitalized
10. In-home isolation
11. Current confirmed cases (isolation + hospital)
12. Change in current confirmed cases from previous day
13. Change in total confirmed cases
14. Recovered
15. Deceased
16. Total positive cases
17. Tests performed
18. Notes in Italian
19. Notes in English

[A.1] Herd Immunity: when most of a population is immune to an infectious disease, this provides indirect protection—or herd immunity (also called herd protection)—to those who are not immune to the disease. *(Source: Johns Hopkins University School of Public Health)*

[A.2] Code for logistic regression model in Approach #3

```
LogisticRegression(C=1.0, class_weight=None, dual=False, fit_intercept=True,
                   intercept_scaling=1, l1_ratio=None, max_iter=100,
                   multi_class='auto', n_jobs=None, penalty='l2',
                   random_state=None, solver='lbfgs', tol=0.0001, verbose=0,
                   warm_start=False)
```