Self-Study

Please solve all exercise questions in groups. I suggest that you meet online with your group on Wednesday at 2:15pm. We have a Zoom meeting on Wednesday 4:45-5:45pm to discuss the solution of the self-study together.
Contact me before the meeting in case you have any questions or problems solving the exercise questions. I will be online from 2:15pm (but you can contact me at any time). You can contact me via email, Skype (writeattony) or Zoom.

Exercise 1: COVID-19 Crisis
We have access to data about 151 countries with at least one confirmed COVID-19 infection from the Johns Hopkins University. The file `countries.csv` contains the country names, continent, total number of confirmed infections, and total number of deaths (as of March 16, 2020). The file `covid19.csv` contains the new daily confirmed infections and deaths from January 23 to March 16, 2020. Figures 1 and 2 visualize the daily confirmed infections and deaths for some selected countries.

Figure 1: Confirmed Infections per Day

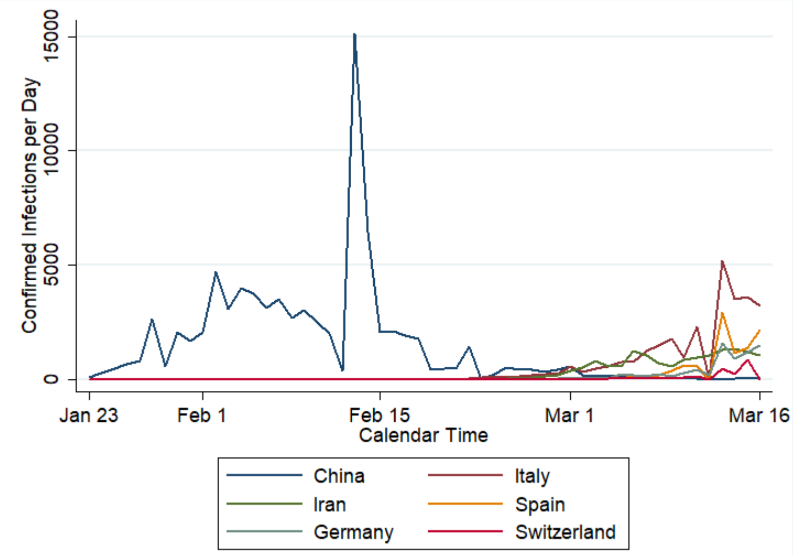
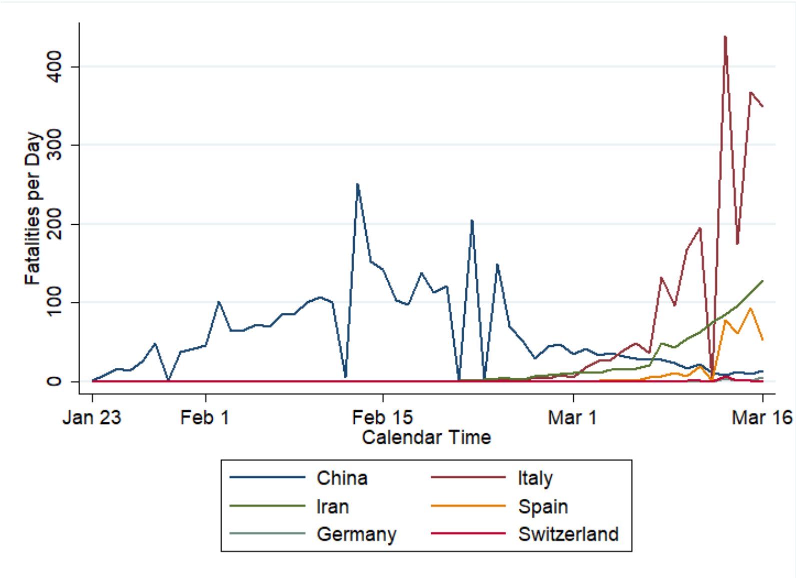


Figure 2: Deaths per Day



1. Load the data files `countries.csv` and `covid19.csv`. Generate a dummy for countries with deaths. Tabulate the number of countries per continent with and without deaths.

2. Regress the confirmed infections on deaths (without an intercept). How many deaths per 1'000 confirmed infections does the model predict? How many deaths does the model predict for Switzerland? How many deaths do we actually have in Switzerland?

3. Generate a variable that ranks the countries based on the total number of confirmed infections. Plot the total confirmed infections (x-axis) against the total deaths (y-axis). Use the `text()` option to label the data points of the seven countries with the most confirmed infections. Add a line with the predicted number of deaths from the regression model above.

4. Cluster the data based on the 108 daily variables in the `covid19.csv` data. How many clusters are optimal? What could be a potential interpretation of the clusters?

5. Use PCA instead of clustering to explore the data. How many principal components are in the data? How many principal components are optimal?

6. Plot the first two principal components. Color the data points by the clusters from the previous exercise.

7. Plot the most important factor loadings of the first two principal components. What could be a potential interpretation of the principal components?

Exercise 2: Wholesale Manager

You manage a wholesale store. The data file `juice.csv` contains orange juice sales (`sales`) and prices (`price`) of different grocery stores that you deliver. Your product range contains three different orange juice brands: `Tropicana`, `Minute Maid`, and `Dominicks`. Some stores advertise/feature specific orange juice brands, which is indicated by the dummy variable `feat`. The data contains also the store ID (`id`).

You deliver new grocery stores. The new stores sent you the file `new_grocery.csv`, which contains the planned prices and advertisements for the different brands. Your job as wholesale manager is to predict the sales of the new grocery stores and deliver the right amount of orange juice.

1. Load the data files. Describe your data cleaning procedure. Are there missing values? Do you transform some variables?

2. How do you partition the data?

3. Asses the performance of Lasso, Ridge, Tree, and Forest estimators in the test sample. Which estimation procedure has the best prediction power? How did you select the tuning parameters? How did you transform covariates?

4. Predict the sales of the new grocery using the estimation procedure with the best prediction power. Save these predictions together with the store ID. Submit the sales predictions and the store ID in one csv-file.